**Summarization of long medical documents with pre-trained transformers**

**Abstract**

Summarization of documents Using modified transformers to improve long medical document summarization is performed with transfer learning. Transformers difficulty with the context window problem affects the model performance. PubMed dataset will be used for the summarization modelling.

**Introduction**

Effective communication and understanding key points have far reaching benefits in domains. Text summarization reduces a long-documents to main ideas and point and helps us to understand these long documents faster. A best example of text summarization that we all are familiar with is news summarization. We have many applications which summarize the long text of news which may take much time to read.

When we think of text summarization in the domain of NLP we have two variants of summarization namely, Extractive summarization and Abstractive summarization. In extractive summarization we select important words and sentences and extract them from the original text thereby create the summary. But more interestingly in abstractive summarization, we can create new summaries without using the exact sentences from the original text without losing any relevant information. Here the entire document is reproduced in a different way. It is just like we read a long article and summarizing with our own words without losing the any relevant information.

**Problem statement**

Even though transformers learn much better long-term dependencies and relations, it still has limited context window problem and this affects the performance of summarization of long documents. So, addressing this problem will be helpful for in several domains.

**Literature review**

**References**

1. Web source- <https://www.machinelearningplus.com/nlp/text-summarization-approaches-nlp-example/>

Extractive summarization – create sentence embeddings and finds relevant sentences and uses the exact sentence in the summary.

Where as in the abstract summarization, will summarize the entire document and reproduce it in a different way, withour loosing the core content. The sentences are not from the original text.

There are many statistical approaches based one word count. We have latent semantic analysis which can be used for extractive summarization. ( it uses singular value decomposition) . Luhn summarization algorithm uses tf-idf for createing the extractive summarization. Kl sum is an another extractive summarization algorithm which is based on word distribution similarity.

It is always better to not use the exact sentences in the original text in the summarization. This is what abstractive summarization . So this is more powerful than just extractive summarization. Now what exactly are transformers. Before going to transformers, lets understand what is rnn , rnn is a king of neural network known as recurrent neural network which is capable of modelling sequential data like time series, natural language etc. Where as LSTM comes with additional memory facility.

Transformers = architecture for sequence to sequence learning. It can handle long term relationship or dependeccy .

Sequence to sequence models are used to convert sequence of type a to sequence of type b. rnn based seq2seq models have attracted a lot of researchers in 2014. The performance of therse seq2seq models was further developed with attention mechanism in 2015. This seq2seq models can be used in many nlp tasks including machine translation, text summarization speech recognition and qa system etc.

There are certain limitation with the seq2seq models with attention.

* Delealing with long -range dependencies is still challenging
* We cant parallelize the model architecture.

Transformers solve this problem of failing with the learning the long term dependencies.

With transformers we are able to implement the abstractive summarization. Transformers provides many pretrained models with good weights.

Transformers have also itown limiatiaations

* Attention can only deal with fixed length text strings. The text has to be split into a certain number of segments or chunks before fed into the system as input
* This led to the context fragmentation
* Transformer architecture can learn longer term dependency. How ever they can’t stretch beyond a certain level due to the use of fixed-length context.

The new sensation in nlp – google’s bert ( bidirectional encoder representations form tranformers)

* Transfer learning attrained a lot of researchers attaraction
* Pretrained deep learning model could be fine tuned for a new task on the imagenet dataset and still give descent results on a relatively small labelled dataset. \
* Bert framework a new language represenatiaon model from google ai, uses pretraining and fine tuning to create state of art models for a wide ragne of tasks. These takss include question anserwering system, ssentiment analysiss and language inforerence.

Literature review

🡪 abstractive summarization of long medical documents with transformers

long document summarizatio using transformers

\* Performed the summarization task using two steps, extractive and abstractive step.

\* used pubmed dataset

\* Results showed that use of pretrained transformers lead to improvements in both extractive and abstractive steps.

\* Transformers have limited-size context windows which limit their ability to perform summarization over long documents; the mixed extractive and abstractive approach attempts to remedy this issue. The models used for extractive and abstractive summarization are trained seperately, and then used sequentially to perform our mixed summarization approach.

\* In the first step, important sentences are extracted from a document through the use of a BERT based extractive summarizer. This summarizer creates sentence-level embeddings for each sentence in the document, which are then passed into either a classifier or cluster algorithm(k means) to identify the best summary sentences.

\* In the second step extracted sentences are fed as input to the BART transformer model. The transformer then creates a summary based off these extracted sentences. Performing the extractive step allows the summary to be conditioned on important sentecnes from throughout the document, despite the limitees context window.

\* Evaluation method - used ROUGE scores

Evaluating extractive text summarization with bertsum

\* Main objective : evaluating BERTSUM using ROUGE metrices.

\* Used - prepared CNN/DailyMail dataset.

\* Note: The goal of extractive text summarization models is to score each sentence in the document to be able to include the most relevant sentences in the summary. In the case of abstractive summarization there is a need for the model to have word generative capabilities given words or context that might not be included in the document. The progress in the extractive text summarization has seen remarkable accuracy thanks to models like BERTSUM which uses fine tuning layers to add document based context from the BERT outputs to more efficient models such as DistllBert which shows relatively similar performance but needs a lot less space and time to run.