**Group 10 Proposal:**

**Abstractive Text Summarization using Deep Learning Techniques**

Group Member: Changyu Liu, Jinda Zhang, Juan Yu, Ye Zhang

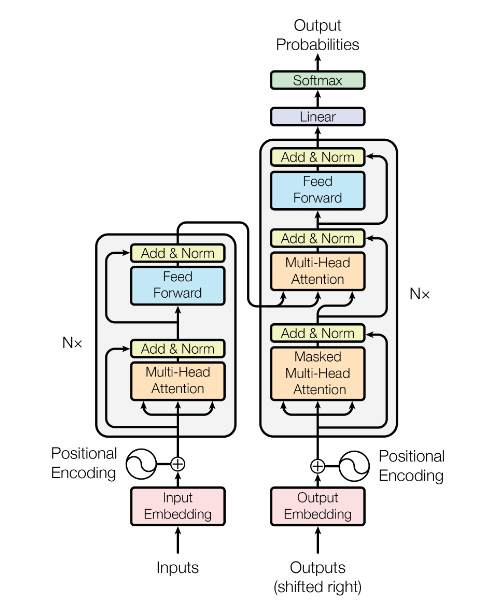
**1. Basic idea**

In the ever-growing digital age, the volume of textual data available online and in print is expanding exponentially. There exists an urgent need to automatically and effectively condense this information into a more manageable and comprehensible form. Our group aims to address the challenge of automating the text summarization process, which refers to the task of producing a concise and coherent summary that retains the most critical information from the source text. Such a solution can have applications in news aggregation, research paper summarization, and much more.

**2. Approach to solution**

There are mainly two types of approaches to text summarization, extractive approaches and abstractive approaches. The extractive approaches rank the sentences from the text to be summarized and put those sentences together as a summary, which means the summary consists of the exact sentences from the original text. In contrast, abstractive approaches interpret the original text and generate new sentences to summarize the text. In this project, we will focus on **abstractive approaches** where more powerful natural language techniques are applied.

Currently, the Transformer architecture is the mainstream architecture for language models, including the extractive text-to-text(summary) models we are interested in.



Source: *Attention Is All You Need*

Following are some of the models we are the most interested in:

**3.1 T5**

T5 [3] is a groundbreaking model, a text-to-text framework that allows us to use the same model, loss function, and hyperparameters on any NLP task.

T5 is a sequence to sequence model, trained to take in a sequence of text and produce another sequence as output. This unified approach allows T5 to be trained on a mixture of tasks, where it learns to convert problem statements into solutions.

Given its design, T5 is highly suited for abstractive text summarization, where the goal is not just to extract relevant parts of the document, but to paraphrase and produce a more concise, coherent summary. T5 can be trained to read the input document and produce a summarized version as its output.

**3.2 Pegasus**

The Pegasus [4] is another Seq2seq model proposed in PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization by Jingqing Zhang, Yao Zhao, Mohammad Saleh and Peter J. Liu. All models are transformer encoder-decoders with 16 layers in each component. Pegasus’ pre-training task is intentionally similar to summarization, Pegasus achieves SOTA summarization performance on all 12 downstream tasks, as measured by ROUGE and human eval.

**3.3 Longformer**

Longformer [5] is a language model based on the transformer architecture, designed to effectively handle lengthy documents. It introduces an innovative global attention mechanism, allowing it to take into account the entire input text, regardless of its length. This is achieved by employing sparse attention patterns, which help in managing computational complexity. Longformer is pretrained on a substantial text corpus and we can fine-tune it for article summarization.

**3. Related work: recent research (2 papers)**

The BERT [1] (Bidirectional Encoder Representations from Transformers) model is built upon the transformer architecture. BERT model is trained bidirectionally, which allows it to grasp language context from both directions. It is pre-trained on vast amounts of text data, which makes it adaptable for various NLP tasks through fine-tuning.

The BART [2] (Bidirectional and Auto-Regressive Transformers) model is a powerful and flexible sequence-to-sequence architecture. It incorporates the strengths of bidirectional pre-training model, enabling it to understand more bidirectional contextual information than GPT. It also absorbs the respective characteristics of GPT's left-to-right decoder, making it more suitable to generate text than BERT.

**4. Assessment methodology**

**4.1 Performance evaluation measures**

BLEU: looking at how many of the tokens in the generated texts are perfectly aligned with the reference text tokens

ROUGE: The approach is very similar to the BLEU score in that we look at different n-grams and compare their occurrences in the generated text and the reference texts. The difference is that with ROUGE we check how many n-grams in the reference text also occur in the generated text.

**4.2 Cross validation strategy**

5-Fold Cross-Validation: 4 fold for training, 1 fold for validation to evaluate the model's performance.

**4.3 Ablation settings**

The pre-trained abstractive models will be fine-tuned, and performances will be compared across models and different combinations of hyper-parameters based on pre-selected evaluation measures.

**5. Reference：**

1. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

2. Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019)

3. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." The Journal of Machine Learning Research 21.1 (2020): 5485-5551.

4. Zhang, Jingqing, et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." *International Conference on Machine Learning*. PMLR, 2020.

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

6. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).