LongFormer

The Long Document Transformer

Niharika Pillanagoyala
810913157

Introduction:

Transformers provides thousands of pretrained models to perform tasks on different modes such as text, vision, and audio. Transformer-based models are unable to process long sequences due to their self-attention operation, which scales quadratically with the sequence length. Models like BERT, RoBERTa have been the state-of-art for a while, the major drawback of these models is that they cannot "attend" to longer sequence. For Example, BERT is limited to max of 512 tokens at a time. To address this limitation, we introduce the "**Longformer – The Long Document Transformer**." with an attention mechanism that scales linearly with sequence length, making it easy to process documents of thousands of tokens or longer.
Longformer, a transformer-based model that is scalable for processing long documents and that makes it easy to perform a wide range of document-level NLP tasks without chunking/shortening the long input and without complex architecture to combine information across these chunks. Below are the types of tasks performed by the LongFormer.

- Question Answering
- Abstractive test summarization
- Document Summarization
- Natural Language Inference
- Text Classification

Question Answering is the task of answering questions (typically reading comprehension questions) but abstaining when presented with a question that cannot be answered based on the provided context. For this task, input representations to separate the question and the document with a special token. The full attention setting will find the mapping between the question and the answer from the document using attention score.
Question answering can be segmented into domain-specific tasks like community question answering and knowledge-based question answering. QA tasks are also divided into domains, depending on where the context for answering the questions can be found:
1. Closed domain/book Question Answering (cdQA): If the answers are localized to one domain, such as medical, legal, etc.
2. Open domain/book Question Answering (odQA): If the answers are in multiple domains, general knowledge questions, or similar.
Dataset: question answering system is SQuAD. (Stanford Questioning Answering Dataset).

Transformer

Transformer is an architecture for transforming one sequence to another one with the help of two parts called Encoder-Decoder. The transformer architecture has become one of the most prominent in natural language processing (NLP) since its introduction in 2017.The primary component of the transformer-based architecture is the use of an **attention mechanism**. Using the attention mechanism, the model learns the relevance for each word in a sentence in relation to other words in the same or other sentences.

A drawback of these models is that the memory requirement and computing time grows quadratically with the input text's length. Transformers cannot process long documents due to its self-attention.

There are also few other approaches for this limitation. Chunking/Shortening the paragraph into 512 words and then aggregating them. But it fails to identify the relation between the words in other chunks which creates information loss due to truncating or cascading.

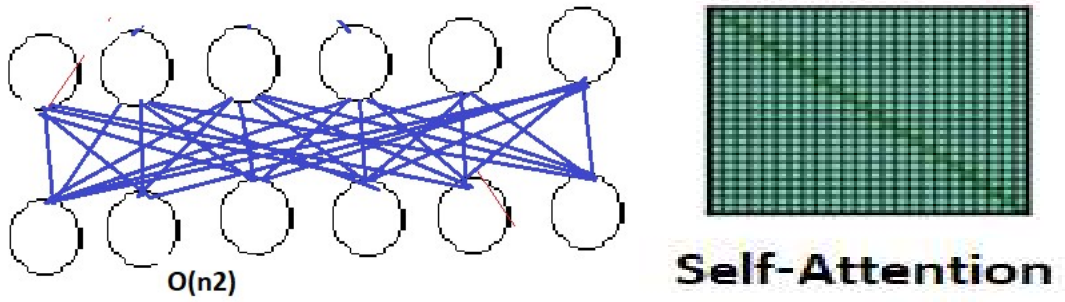
Longformer:

Longformer is a modified Transformer architecture. Longformer uses encoder-decoder architecture like the original Transformer model. It uses an attention pattern that scales linearly with sequence length, making it easy to process documents of thousands of tokens or longer. The attention mechanism is a drop-in replacement for the standard self-attention and combines a local windowed attention with a task motivated global attention.

Encoder-Decoder Architecture - The Encoder-Decoder architecture consists of two neural networks: an encoder that takes in some input sequence and encodes it to a fixed-length lower-dimensional vector representation of the text; and a decoder that takes as input the vector from the encoder and reconstructs the text sequence as output. The encoder and decoder are trained jointly, where the goal is to encode and reconstruct a target sequence as closely as possible.

Attention Mechanism: Transformer VS LongFormer

Transformer Self-Attention: General Classic transformers use self-attention. In Self-Attention each token in layer 1 will attend to each token in layer 2. Suppose if we have n sequence of words as an input and each word is considered as a token. Each token passes through each other tokens in another layer and has n outputs. So overall calculation comes as n2 which increases memory usage by 4 times also increases the computation.



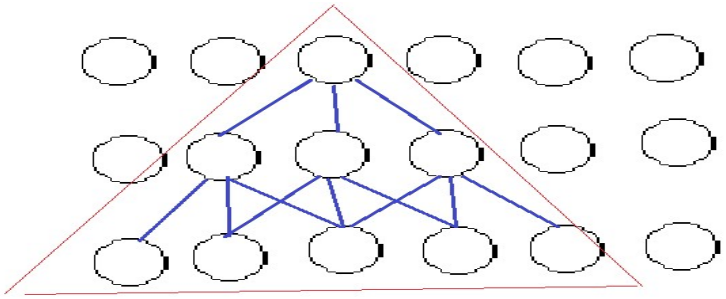
The self-attention layer will learn. It will essentially learn a contextualized meaning for each word in the input function. It simply calculates a similarity score between each word and the rest of the words. That is then normalized (with SoftMax) to become an attention score, and each word is then weighted by its similarities to other words.

Then each of those "contextualized-meaning embeddings" are then put through the same 2 layer, fully connected feed-forward network - which has an output of the same size (512), with a much larger hidden layer. The role and purpose are to process the output from one attention layer in a way to better fit the input for the next attention layer.

LongFormer Attention: To Overcome long sequence issues, the LongFormer combines several attention patterns sliding window attention, Dilated sliding window and Global+sliding window.

Sliding Window:

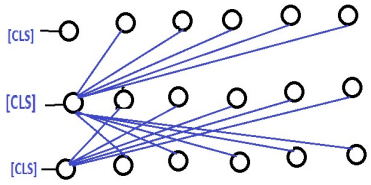
Attention pattern uses a fixed-size window attention surrounding each token and its neighboring tokens (i+1, i, i-1). Using multiple stacked layers of such windowed attention results in a large receptive field, where top layers have access to all input tokens and have the capacity to build representations that merge information across the entire input. Given fixed window size w, each token attends 1/2w token on each size. The computation complexity is O(w*n) and when window size is constant O(w*n) becomes O(n) which scales linearly with the input sequence length and reduces the computational function.

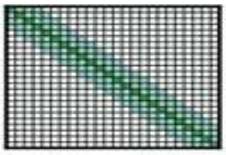


Global + sliding window:

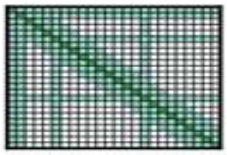
Memory Extension is a common method utilized by several models to allow for a limited number of tokens to attend to every other token in the sequence. This attention model is commonly referred to as global attention. These special global attention tokens can either be learned or assigned manually. The separation token <s> often assigned with global attention tasks to allow better retention between the context and the question. For classification, the model aggregates the representation of the whole sequence into a special token ([CLS]. Since the number of such tokens is small relative to and independent of n the complexity of the combined local and global attention is still O(n).

The selected tokens ae very important and these are defined by the developers. These tokens are connected to all the nodes. Every single path is the max length of 2 because if you want to connect to nodes 1 and 5 send info to global node and this global node will send info to the target node which achieves linear complexity in terms of memory computes.





Sliding Window



Global+Sliding window

For Question Answering (QA), the text's answer needs to map to the question asked. For longer sentences, not having global attention in certain instances severely reduces the performance of the model performance. By allowing a certain number of tokens k, such as the starting-, separating and end tokens, to have global attention, the mapping for these tasks drastically improves while keeping the number of tokens with dense self-attention to a minimum.

Training:

We trained model on the largest window size and sequence length that fits in a modern GPU memory, and we found that the model needs many gradient updates to learn the local context first, before learning to utilize longer context. We have assumed a staged training procedure where we increase the attention window size and sequence length across multiple training phases.
Trained the model over 5 total phases with starting sequence length of 2,048 and ending sequence length of 23,040 on the last phase.
In the first phase we start with a short sequence length and window size, then on each subsequent phase, we double the window size and the sequence length, and halve the learning rate. This makes training fast, while keeping the slow part (longest sequences and window sizes) to the end.

Results:

Below is the Wikipedia results when you ask a question "which name is also used to describe Amazon rain"

Amazon rainforest

From Wikipedia, the free encyclopedia

Coordinates: 3°S 60°W

"The Amazon" and "Amazonia" redirect here. For the river, see Amazon River. For other uses, see Amazon and Amazonia (disambiguation).

The **Amazon rainforest**, alternatively, the **Amazon jungle**[a] or **Amazonia**, is a moist broadleaf tropical rainforest in the Amazon biome that covers most of the Amazon basin of South America. This basin encompasses 7,000,000 km^2 (2,700,000 sq mi), of which 5,500,000 km^2 (2,100,000 sq mi) are covered by the rainforest. This region includes territory belonging to nine nations and 3,344 formally acknowledged indigenous territories.

The majority of the forest is contained within Brazil, with 60% of the rainforest, followed by Peru with 13%, Colombia with 10%, and with minor amounts in Bolivia, Ecuador, French Guiana, Guyana, Suriname, and Venezuela. Four nations have "Amazonas" as the name of one of their first-level administrative regions, and France uses the name "Guiana Amazonian Park" for its rainforest protected area. The Amazon represents over half of the planet's remaining rainforests,[2] and comprises the largest and most biodiverse tract of tropical rainforest in the world, with an estimated 390 billion individual trees divided into 16,000 species.[3]

More than 30 million people of 350 different ethnic groups live in the Amazon, which are subdivided into 9 different national political systems and 3,344 formally acknowledged indigenous territories. Indigenous peoples make up 9% of the total population with 60 of the groups remaining largely isolated.[4]

Contents [hide]

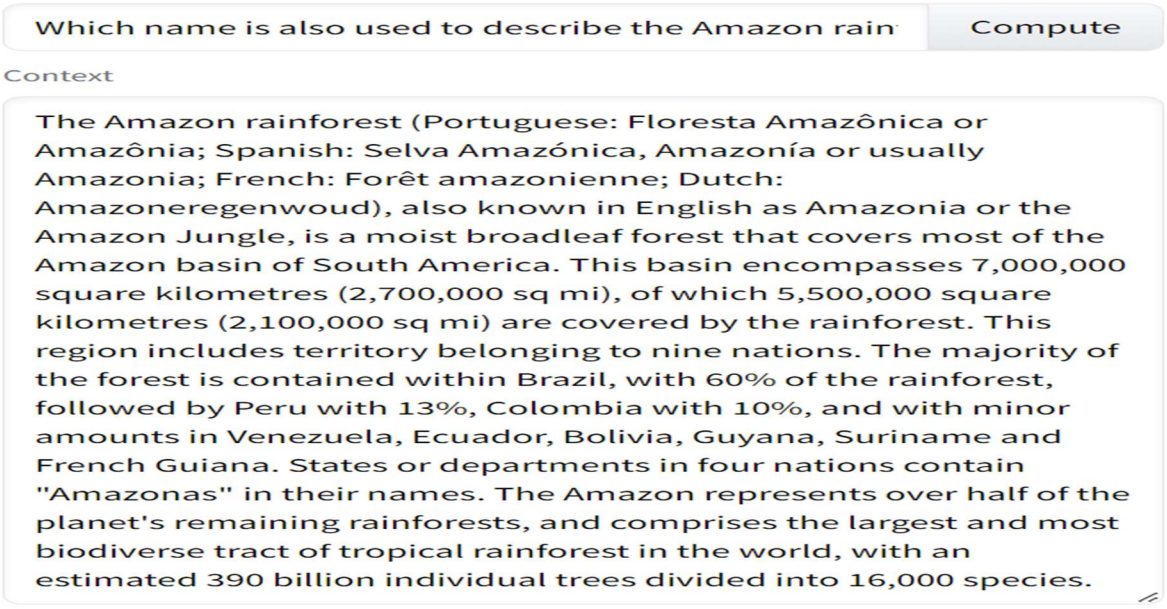
1 Etymology
2 History
3 Geography
  3.1 Location
  3.2 Natural
    3.2.1 Sahara Desert dust windblown to the Amazon

Amazon rainforest
Portuguese: *Floresta amazônica*
Spanish: *Selva amazónica*

Amazon rainforest on the Urubu River

Map

These are the results obtained by the LongFormer:



Drawback and Future Work:

Drawback of these models is that research mainly has been constricted to English, and most efficient transformers need to be trained from scratch using long-context datasets.

This is very computationally expensive, and such datasets may not be available in other languages, especially low-resource languages. Research of transformer-based models has primarily been made on for the English language or other so-called high-resource languages. Among the over 7000 languages globally, ten of these comprises 76.9% of the internet presence and English alone for 25.9%.1 A language such as Swedish for instance is considered a low-resource language because of the limited number of articles on the internet, few speakers and limited number of high-quality datasets

Conclusion:

Longformer can process documents without truncating or chunking, allowing to adopt a much simpler approach that concatenates the available context and processes in a single pass.

Reference Links:

The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)

2004.05150.pdf (arxiv.org)

Question Answering System using Transformer | Neurond AI | Medium