BigBird and Clinical-BigBird

FADIL MIR

Motivation

- Transformers-based models, such as BERT, have been one of the most successful deep learning models for NLP.
- One of their core limitations is the quadratic dependency (mainly in terms of memory) on the sequence length due to their full attention mechanism.
- To remedy this, BigBird model was proposed, that uses a sparse attention mechanism that reduces this quadratic dependency to linear.

Related Work

- There have been a number of attempts, that were aimed at alleviating the quadratic dependency of Transformers.

- SpanBERT, ORQA, REALM have achieved strong performance for different tasks. These models used mechanisms to select a smaller subset of relevant contexts to feed into the transformer. However, these methods often require significant engineering efforts and are hard to train.

- Several other models have been developed which used approaches that do not require full attention.

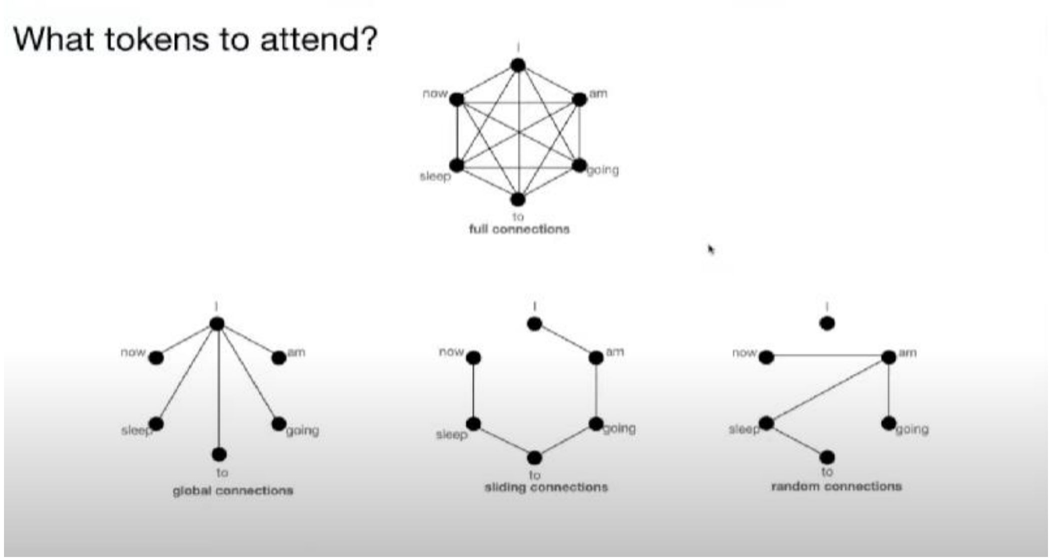
Architecture

- BigBird runs on sparse attention mechanism that allows it to overcome the quadratic dependency of BERT.
- In particular BigBird consists of three main parts:

i. A set of g global tokens attending on all parts of the sequence.

ii. All tokens attending to a set of w local neighboring tokens.

iii. All tokens attending to a set of r random tokens.



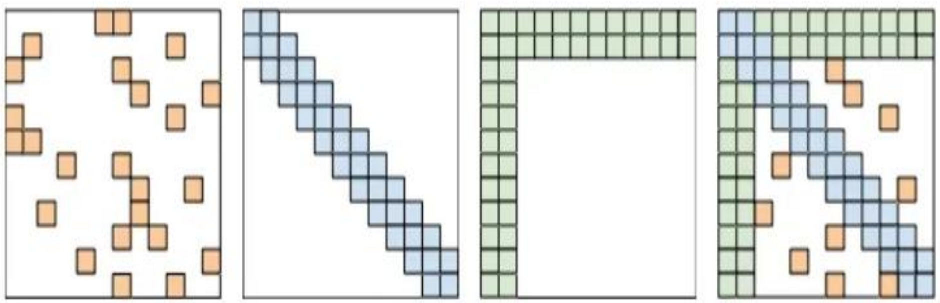
• Attention Mechanism:

- BigBird runs on sparse attention mechanism that makes it possible to have a linear complexity.. It's attention mechanism is a combination of:

i. Random Attention

ii. Window Attention

iii. Global Attention



(a) Random attention (b) Window attention (c) Global Attention (d) BigBird

- ## Maximum input size:

In BERT, the maximum input size is 512 tokens because of quadratic nature of it's complexity in terms of computation.

BigBird can process sequences of length 8x more than BERT(i.e. 4096 tokens)

- ## Content Fragmentation:

In BERT, content fragmentation is present because longer sequences have to be broken into smaller segments.

BigBird overcomes the problem of content fragmentation.

Performance Comparison

- **Question Answering Task(QA):**
- BigBird was found to be performing better than models like RoBERTa, Longformer, SpanBERT on various QA datasets like HotpotQA, Natural Questions, TriviaQA, and WikiHop.

- **Classification:**
- BigBird performs better in document classification and various GLUE tasks. It improves state-of-the-art for Arxiv dataset by about 5% points. On Patents dataset, there is improvement over using simple BERT/RoBERTa.

Performance on Clinical data

- The Pre-trained BigBird model was not found to be performing well on clinical dataset for sentiment analysis tasks.

- When used for a custom clinical dataset, it was observed that that BigBird was incorrectly labelling various samples.

```
[ ]  Prediction

     array([0, 0, 0, 0, 0, 0, 0, 0])

[ ]  from sklearn.metrics import accuracy_score , confusion_matrix,ConfusionMatrixDisplay

     acc=accuracy_score(Prediction,truth)
     acc

     0.5
```

Clinical-BigBird

- Inspired by the success of long sequence transformer models like Longformer and BigBird, domain enriched language models were introduced.
- One such model is the Clinical-BigBird, which is pre-trained from large-scale clinical corpora.
- It has achieved state-of-the-art results when performed on clinical named entity recognition and natural language inference tasks.

Related Work

- Transformer-based models, especially BERT, can be enriched with clinical and biomedical knowledge through pre-training on large-scale clinical and biomedical corpora.
- These domain-enriched models, e.g. BioBERT pretrained on biomedical publications and ClinicalBERT pre-trained on clinical narratives, set the state-of-the-arts when down-stream applied to clinical and biomedical NLP tasks.
- However, these models were built on the basic BERT architecture, which has a limitation of 512 tokens in the input sequence length.

Performance Comparison

- Clinical-BigBird has been found to outperform models like BERT, BioBERT and ClinicalBERT on various clinical Question Answering datasets.
- Similarly, it has better on various named entity recognition tasks and document classification datasets.

Performance on Custom dataset

- The Pre-trained Clinical-BigBird model was found to be performing well on clinical dataset for sentiment analysis tasks.

```
[ ]  Prediction

array([0, 1, 0, 0, 1, 1, 0, 0])

[ ]  truth = df.iloc[:,1].values

[ ]  truth

array([0., 1., 0., 1., 1., 1., 0., 0.])

[ ]  from sklearn.metrics import accuracy_score , confusion_matrix,ConfusionMatrixDisplay

[ ]  acc=accuracy_score(Prediction,truth)
     acc

0.875
```

THANK YOU