Rishu Ray Singh
2023 AFB 2080

Page No.:
Date:

Hands on 5

Transformers

Task:-
To Develop a transformer architecture with given configurations from Scratch for language translation.

Dataset :-
The dataset, contains two data columns, one column has english words/sentences and other one has French words/sentences.

Configurations:-

a) A single attention head with 3 encorders and decoders respectively.

b) Multi attention heads of 8 with 3 encoders and decoders respectively.

c) Multi attention heads of 32 with 3 encoders and decoders respectively.

d) Multi attention heads of 8 with 5 encoders and decoders respectively.

a) configuration - 1

This configuration relies on single mechanism to weigh the importance of different words. It has slighly struggled to capture complex dependencies

But, by increasing number of encoder and decoder layer, it has gained capacity to learn hierarchical representances of the input and output sequences.

Results:-   Training-loss = 0.21077

Plot-1   val.-accura = 88.73 %

Test_accura = 88.60 %

Test_BLEU-Score = 46.94 %

b) configuration - 2

Increasing the multiple attention heads allows the model to attend to different part of input sentences simultaneously.

But, it has increased computational complexity.

Here, on comparison with configuration '1', It takes more time to train and on average, it gives similar results as of configuration '1'.

Results:-

Plot-2

Trainging_loss =	0.17920
val- accuracy =	88.70 %
Test- accuracy =	88.60 %
Test- BLEU score =	46.60

c) Configuration - 3
Multi head attention with 32, should further enhances the model ability to capture diverse linguistic features and dependencies.

But with given configuration of running '200' epochs has led my model to overfit and hence my model val-accuracy decreases in comparison to configuration one and two. which should not be the ideal case.

Results :-

Plot-3

Training loss =	0.21347
val- accuracy =	88.55 %
Test - accuracy =	88.32 %
Test_BLEU Score =	43.37 %

d) Configuration - 4
with 8 multi head attentions and 5 each encoder and decoder blocks has basically enhances the model's ability to capture complex patterns and relationships in the data. The increased depth allows for more expressive representation.

On comparising it with configuration '2' it has given better results with respect to val-accuracy, test-accuracy. it captures the data complexity more accurately than config-2.

Results:- Training - loss = 0.,14989

plot-4 Val - accuracy = 89.91 %

Test - accuracy = 89.77 %

Test - BLEU - Score = 46.35

Strengths:-

-) configuration 2 and configuration 4, config: has / ~~likely~~ outperform configuration 1 due to use of multi attention head, enabling better capturing of contextual information.

•) configuration 4 with increased encoder and decoder, has excel in capturing complex linguistic structures and nuances, especially in longer sentences.

weakness:-

•) configuration 3, with 32 attention head had suffered from increased computational complexity, leading to longer training time and higher resources allocation.

•) configuration 4, while potentially more accurate has also suffer from increased complexity making it harder to interpret and deeply deploy in resource - constrained setup.