Investigating the Impact of Parameter Variations of Transformer Models on Sentiment Classification

Peng Chen *

Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Champaign, America

* Corresponding Author Email: pengc5@illinois.edu

Abstract. With the development of the Internet, an increasing number of applications and websites appearing on the Internet allow movie viewers to add comments for movies. If people want to know what per cent of comments are positive or negative, it will cost a large amount human resources and time to check each comment. However, with the help of the transformer model, it will save a large number of human resources and time to finish sentiment classification for long movie comments. The dataset 'IMDb' used to train the transformer model is a large Movie Review Dataset for binary sentiment classification of movie reviews. Furthermore, since sentiment classification for movie comments does not require the decoder in the transformer model to predict the next token, the transformer model only need to preserve the part of Positional Encoding, Encoder and Multi-head self-attention mechanism. This paper will investigate how three parameters (the number of layers in encoder, the amount of heads in multi-head self-attention, expansion factor in position-wise fully connected feed-forward network) affect the performance of the transformer model and which set of parameters could allow the transformer to have the best performance. After researching on three parameters, the transformer model used to do sentiment classification for movie comments has the best performance when there are sixteen heads in multi-head self-attention mechanism, four layers in Encoder, and expansion factor in feed-forward network which is position-wise fully connected is four.

Keywords: Transformer, Sentiment classification, deep learning.

1. Introduction

Watching movies is one of the most important ways of entertainment for people in modern society. Furthermore, as the development of Internet, there are an increasing number of websites and application software offering the service of adding comments for movies to movie viewers appearing, and this is very meaningful since comments from movie viewers who have watched these movies will give valuable advice to people who would like to watch them later. However, sometimes, there are too many long comments for some movies, and people who aim to watch these movies are inpatient to spend time reading these comments, but luckily, current Artificial Intelligence (AI) model can help these people finish the task of sentiment classification for these long movie comments quickly due to their excellent performance in many fields such as engineering, healthcare and biometrics [1-5], which means people can know whether the comment is negative or positive in a very short time. As a result, using AI model to do the task of sentiment classification for long movie comments should receive more attention.

In recent several years, an increasing number of strong AI models appear, and they perform very well in Natural Language Processing (NLP). Also, Lin et al. mentioned that it is a significant task to classify the sentiment of user comments on a website within Natural Language Processing (NLP) [6]. Traditional machine learning models and neural networks can all be used to classify the sentiment of user comments but when the transformer based on attention mechanisms appears, it can be said that AI model changed dramatically. Transformer is a very strong AI model, which performs very well within NLP, and it can be used to do the task of text generation, translation, sentiment analysis for text. Compared with traditional machine learning models and some neural networks, attention mechanisms allow transformer to have its own advantages especially when processing and understanding long text, and just the paper 'Attention Is All You Need' mentioned that for compelling

sequence modeling and transduction models in various tasks, attention mechanisms have become an integral part, which allows constructing relationship among tokens in sequences without the effect of distance among tokens [7]. so due to its strong ability of understanding long text, the most well-known pre-training AI models currently like GPT, BERT and so on are all based on the architecture of transformer, and with the optimization and addition of both parameters and modules for the basic architecture of transformer, these AI models are powerful enough to help people solve most problems in life, working and studying. In addition, many researchers pay attention to optimize the architecture of transformer and compare different models previously, and there are not too many people researching on basic parameters for the transformer but these basic parameters are very significant for transformer. Therefore, in the later part of the paper, it is mainly about the optimization of basic parameters of transformer to make the transformer perform better in sentiment analysis of movie comments.

For the architecture of transformer, this paper decided to use its basic architecture since it will make the model be trained faster using GPU and be easier to observe fluctuations for the accuracy of the model when changing basic parameter. Also, this paper will delete the part of decoder of the transformer since sentiment analysis of movie comments does not need a decoder to predict the next token. For basic parameters, this study will pay attention to the number of heads, the number of layers in encoder, and so on. Also, the dataset that this paper chose is called 'IMDb', which is a large dataset used to finish binary sentiment classification of movie reviews.

2. method

2.1. Dataset preparation

The dataset 'IMDb' used in this study is from hugging face, and it is created by Standford NLP [8]. "IMDb" is a large Movie Review Dataset used to finish binary sentiment classification of movie reviews, which contains a group of 25000 movie reviews which is used for training and 25000 movie reviews which is used for testing. Binary sentiment classification of movie reviews is separated into negative or positive by the transformer.

Before constructing the transformer model and training it, preprocessing the dataset is one required task. Before preprocessing the dataset, it should load the 'IMDb' dataset. Then loading BERT tokenizer from 'transformers' library could finish the task of embedding which allows texts to be converted into vectors. The next part is creating dataloaders in order to train and test the transformer. For the architecture of the transformer model, it can be separated into Positional Encoding, Encoder, Decoder and Attention.

2.2. Transformer-based prediction

Since sentiment classification of movie reviews does not need the transformer to predict the next token, the part of Decoder should be deleted, and all the other parts should be kept. For Encoder, there are several identical layers (the number of layers is six by default), and for each layer, it contains two sub-layers: multi-head self-attention and feed-forward network which is position-wise fully connected. Also, a residual connection exists around two sub-layers, which is followed by layer normalization. Feed-forward network which is position-wise fully connected consists of two linear transformations and a ReLU activation, which can be shown as$FFN(x) = max(0, \mathrm{x} \cdot \mathrm{W1} + \mathrm{b1}) \cdot \mathrm{W2} + \mathrm{b2}$. Furthermore, multi-head self-attention lets the transformer model have a good performance in analyzing long text, and it is the core of the transformer, which is composed of several Scaled Dot-Product Attention ((the number of attention layers are six by default), linear projections etc.

The paper mentioned that multi-head attention achieves the function that making the model successfully deal with information which is from different representation subspaces at different positions together [7]. Scaled Dot-Product Attention follows the formula that $Attention(Q, K, V) =$

$softmax(\frac{Q \cdot K^T}{\sqrt{dk}}) \cdot V$ where Q is queries, K is keys, V is values, and softmax is a softmax function. For Positional Encoding, it is used to record the order of the text sentence, and two math functions are applied: $\mathrm{PE}(\mathrm{pos}, 2\mathrm{i}) = \sin(\mathrm{pos}/10000^{2\mathrm{i}/\mathrm{dmodel}})$ and $\mathrm{PE}(\mathrm{pos}, 2\mathrm{i}+1) = \cos(\mathrm{pos}/10000^{2\mathrm{i}/\mathrm{dmodel}})$ where dmodel is the dimensionality.

2.3. Implementation details

In addition, there are totally eight parameters in the transformer model: size of vocabulary, dimensionality of token embeddings, the amount of layers in encoder, the amount of heads in the part of multi-head self-attention, expansion factor in feed-forward network which is position-wise fully connected, dropout rate, maximum length of sequences, and the number of classified classes, and their initial value are set to 3000, 512, 6, 8, 4, 0.1, 512 and 2. After constructing the transformer by PyTorch, training the transformer with training dataset is also very significant. The optimizer chosen is Adam [9, 10], and the learning rate for it is$3 * e^{-5}$. Furthermore, the initial number of epochs are set to three, and in each epoch, the train loader which is already processed in dataset preprocessing will be iterated, and each iteration is through one batch of data in train loader and calculates the loss.

After training the transformer model, the next step is to evaluate the trained model. In this study, the performance of the transformer model is shown by the result of accuracy and loss. This paper mainly researches how parameters of the transformer affect the performance and what values of basic parameters could allow the transformer to have the best performance in binary sentiment classification of movie reviews.

This paper mainly investigates three parameters: the number of layers in encoder, the number of heads in the part of multi-head self-attention, expansion factor in feed-forward network which is position-wise fully connected due to their importance. These three parameters are crucial for the transformer which is used to do the binary sentiment classification of movie reviews so researching on them must bring large enhancement of performance to the transformer model.

3. Results and discussion

After training the transformer model with different sets of parameters and evaluating them fully, the performance of the transformer used to do the task of sentiment classification for long movie comments with different sets of parameters can be presented by results of these evaluations. To analyze and evaluate the transformer model fully, the experiment uses two evaluation metrics: Loss and Accuracy.

The number of heads in the part of multi-head self-attention has a significant effect shown in Figure 1 and Figure 2 to the performance of the transformer model. In the experiment, there are four parameter sets which only modify the number of heads in the part of multi-head self-attention testing how many heads is the best. Below is results of evaluating the transformer model's performance with different numbers of heads in two evaluation metrics:

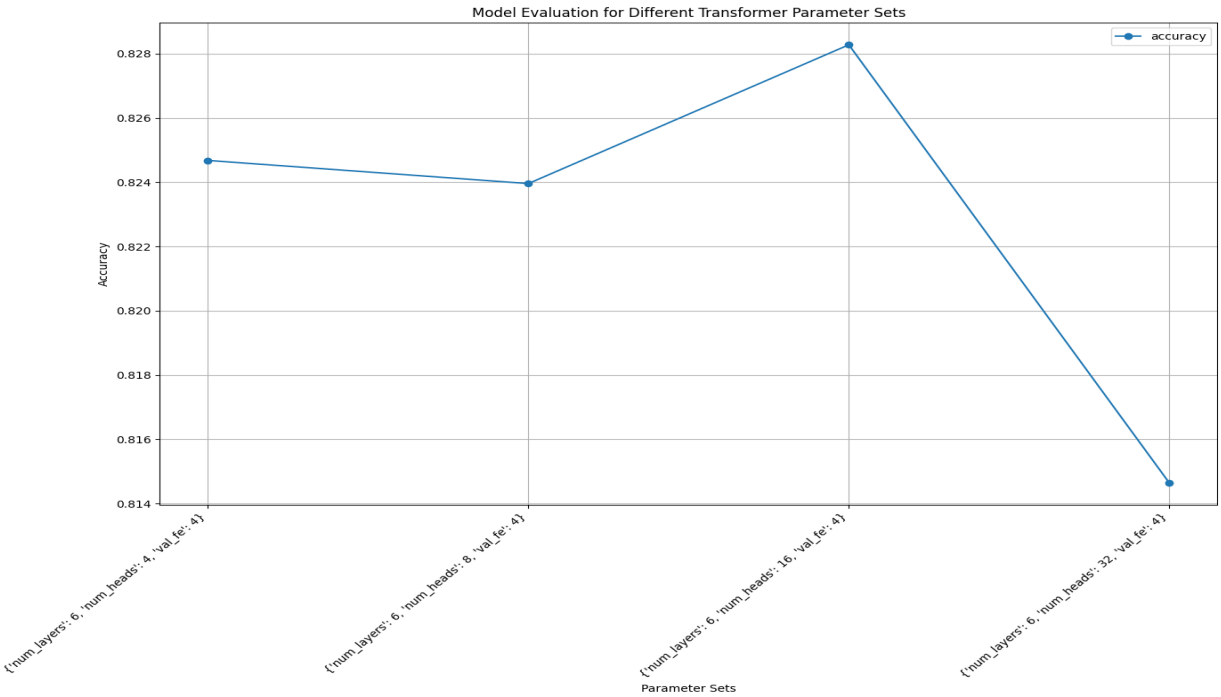

Figure 1. Accuracy of Different number of heads for the transformer (Photo/Picture credit: Original).

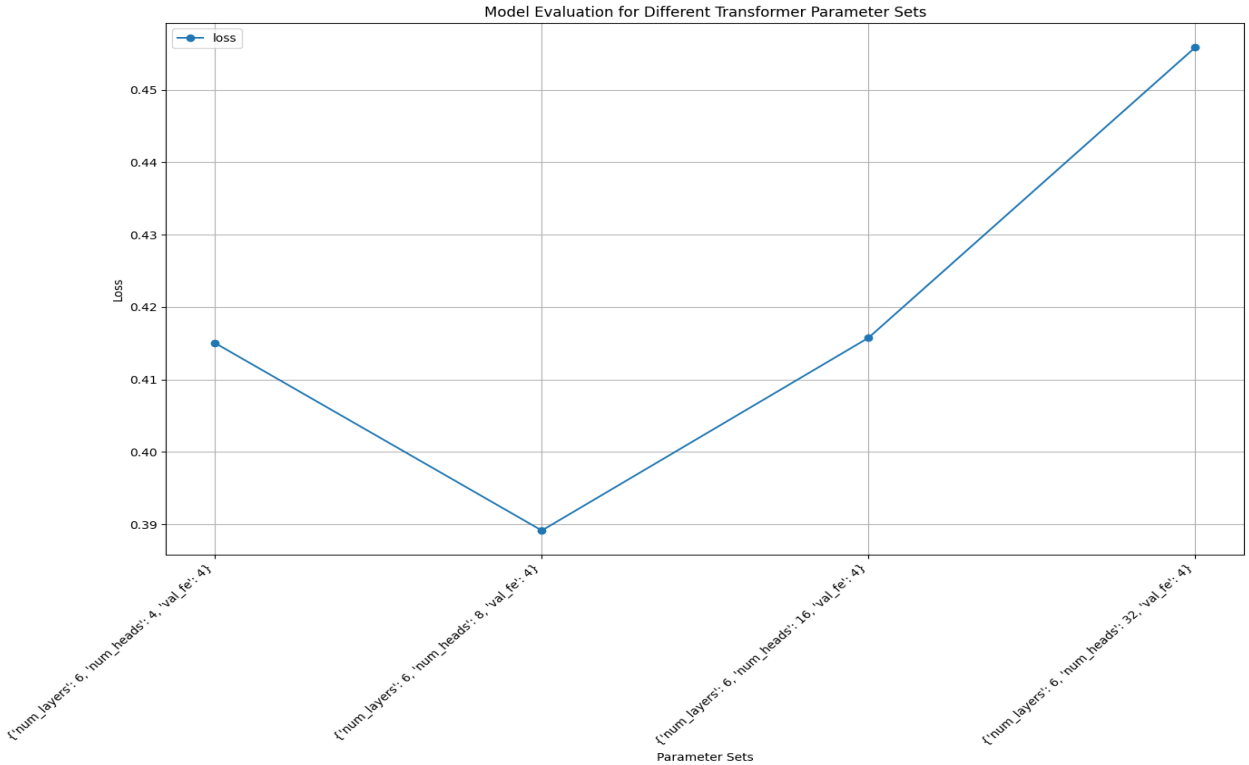

Figure 2. Loss of Different number of heads for the transformer (Photo/Picture credit: Original).

By comparing results with different number of heads, the transformer model performs the best when the number of heads in the part of multi-head self-attention is sixteen. Furthermore, different numbers of layers in encoder bring huge influence on the model's performance. Figure 3 and Figure 4 are results of evaluating the transformer model's performance with different numbers of layers in two evaluation metrics:

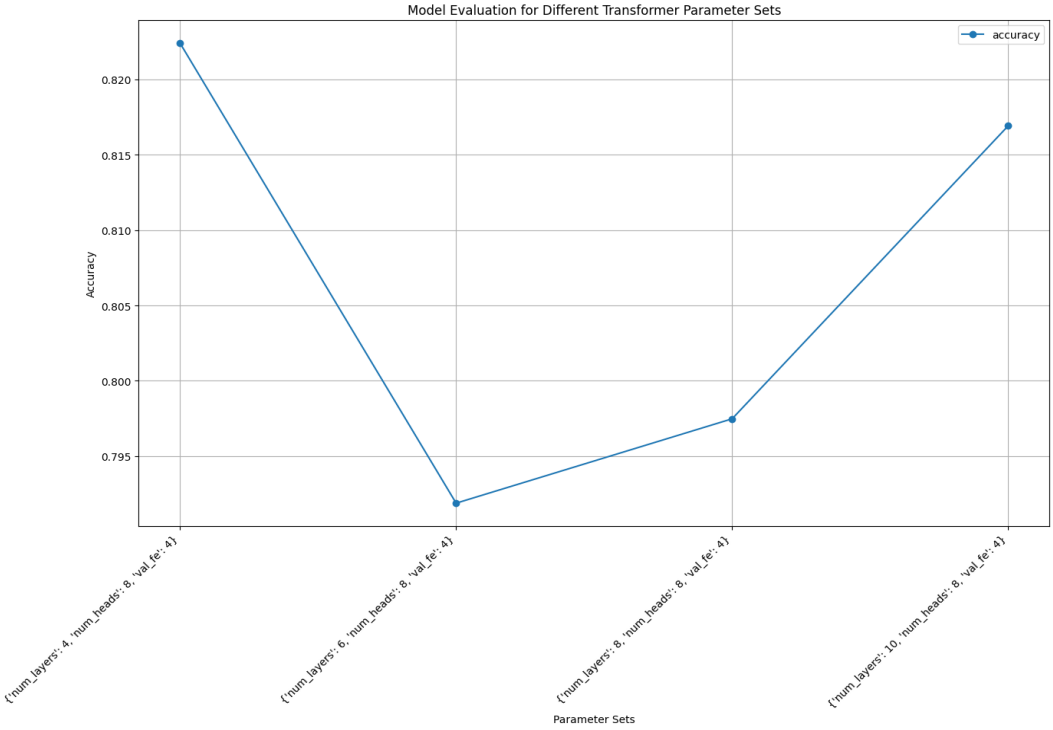

Figure 3. Accuracy of Different number of layers for the transformer (Photo/Picture credit: Original).

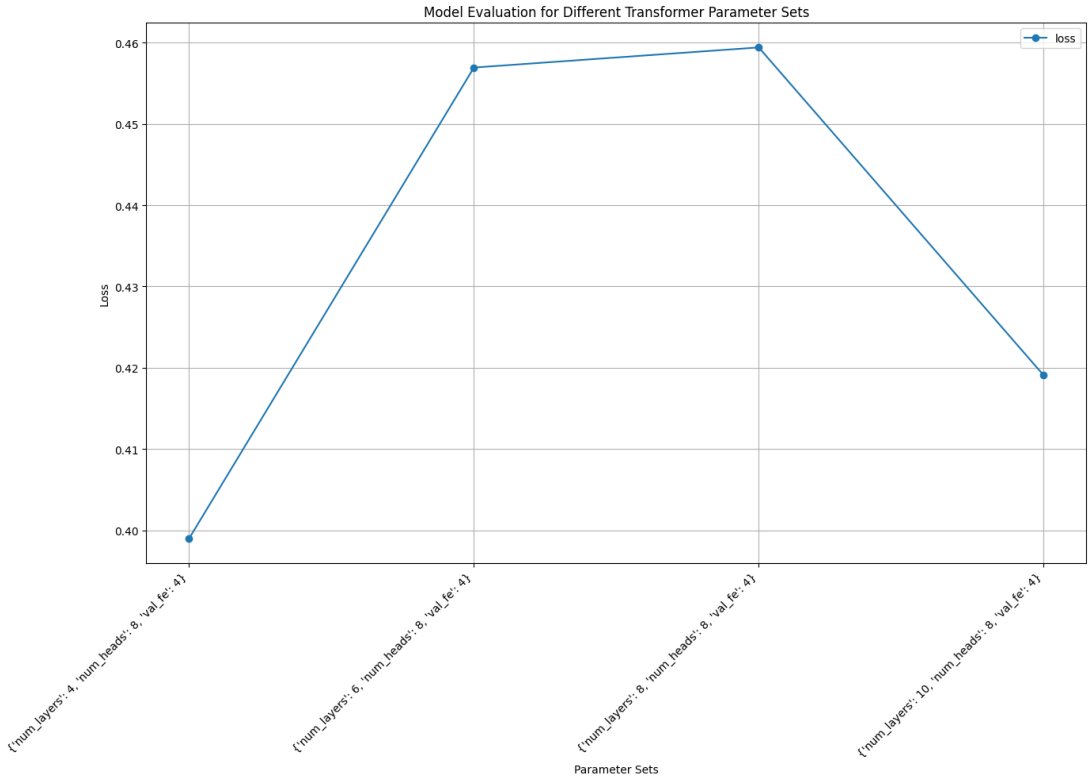

Figure 4. Loss of Different number of layers for the transformer (Photo/Picture credit: Original).

From results with different number of layers in Encoder, it seems that the transformer model performs the best when the number of layers in Encoder is four. Finally, different expansion factors in feed-forward networks which are position-wise fully connected have a huge effect to the model's performance since FFN is an important part for the transformer model. Figure 5 and Figure 6 are results of evaluating the transformer model's performance with different expansion factors in two evaluation metrics:

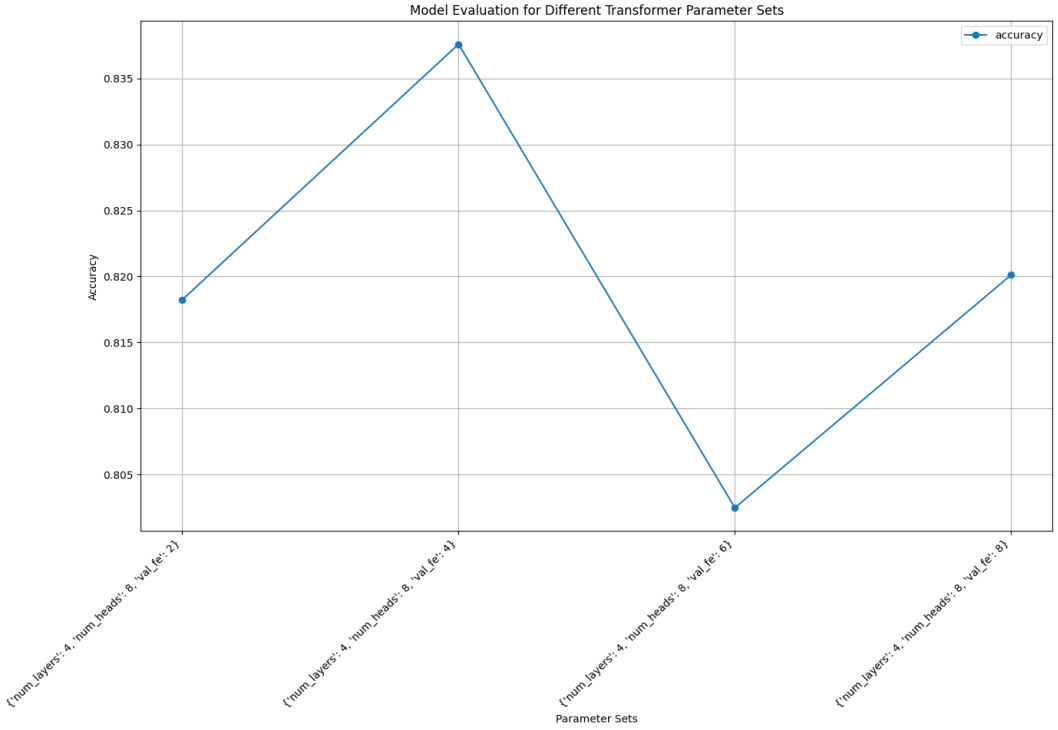

Figure 5. Accuracy of Different expansion factors for the transformer (Photo/Picture credit: Original).

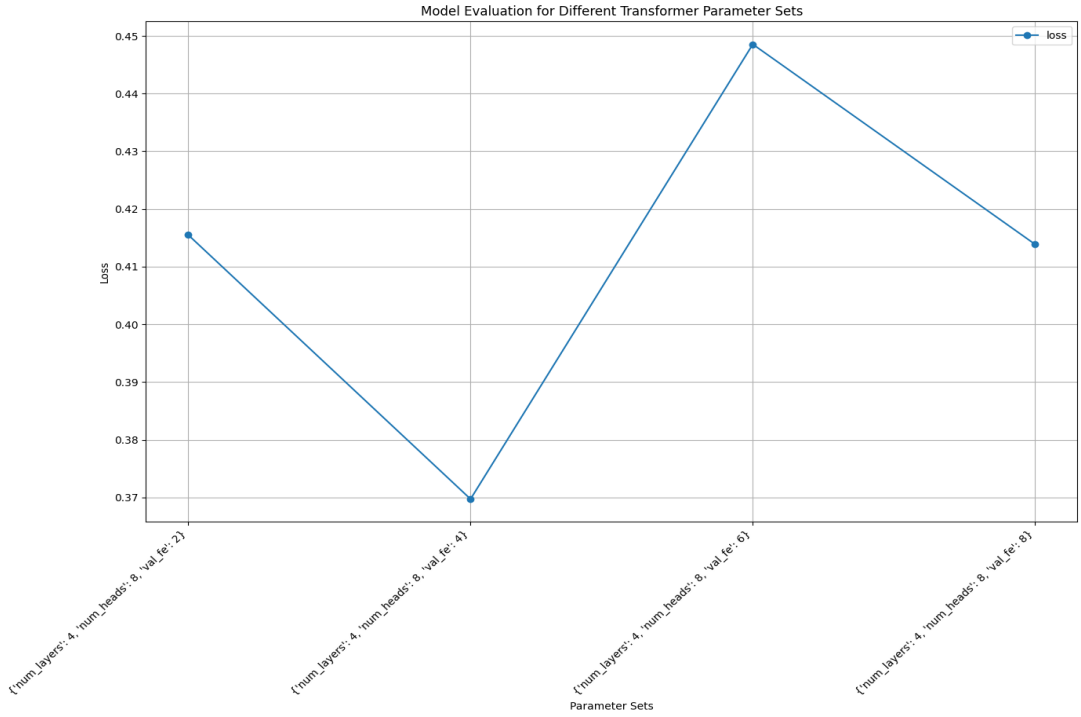

Figure 6. Loss of Different expansion factors for the transformer (Photo/Picture credit: Original).

When analyzing results with different expansion factors in position-wise fully connected feed-forward networks, the transformer model performs the best when expansion factor in feed-forward networks which is position-wise fully connected is four.

When the number of heads in the part of multi-head self-attention is larger than sixteen, the model maybe overfitting so that it will not perform well, and smaller number of heads will make the model not complex enough. So, sixteen heads in the part of multi-head self-attention mechanism make the transformer model best get dependencies among different words of sequences. For the number of

layers in Encoder, just like the amount of heads, too many or too few layers will cause overfitting and the problem of not complex enough, and four layers in Encoder make the transformer model trained by 'IMDb' dataset perform the best. For expansion factor in feed-forward networks which is position-wise fully connected, it can help FFN have more complex linear transformation to enhance the transformer model's performance. And large expansion factor will cause overfitting; however, too small expansion factor is not effective to optimize FFN. Therefore, when expansion factor in feed-forward networks which is position-wise fully connected is four, the model's performance is the best.

4. Conclusion

In conclusion, this paper studied which set of parameters in transformer model perform best when it is used to do sentiment classification for long movie comments. And the transformer model is trained by the dataset 'IMDb' which is a large Movie Review Dataset used to finish binary sentiment classification of movie reviews. Furthermore, the amount of layers in encoder, the amount of heads in multi-head self-attention and expansion factor in feed-forward network which is position-wise fully connected are three parameters which are researched in this paper, and after fully researching on different sets of these three parameters , when the amount of layers in encoder, the amount of heads in multi-head self-attention and expansion factor in feed-forward network which is position-wise fully connected are four, sixteen and four, the transformer model performs the best in sentiment classification of movie reviews. In the future, the further study plans to research on set of parameters for the transformer model trained with more different datasets, which make the set of parameters more comprehensive.

References

[1] Yuhang Qiu, Yunze Hui, Pengxiang Zhao, Cheng-Hao Cai, Baiqian Dai, Jinxiao Dou, Sankar Bhattacharya, Jianglong Yu. 2024. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. Energy, 294: 130866.

[2] Binyang Song, Rui Zhou, Faez Ahmed. 2024. Multi-modal machine learning in engineering design: A review and future directions. Journal of Computing and Information Science in Engineering, 24(1): 010801.

[3] Kashif Shaheed, Piotr Szczuko, Munish Kumar, Imran Qureshi, Qaisar Abbas, Ihsan Ullah. 2024. Deep learning techniques for biometric security: A systematic review of presentation attack detection systems. Engineering Applications of Artificial Intelligence, 129: 107569.

[4] Yuhang Qiu, Jiping Wang, Zhe Jin, Honghui Chen, Mingliang Zhang, Liquan Guo. 2022. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72: 103323.

[5] Allard W. Olthof, Peter M.A. van Ooijen, Ludo J. Cornelissen. 2021. Deep learning-based natural language processing in radiology: the impact of report complexity, disease prevalence, dataset size, and algorithm type on model performance. Journal of Medical Systems, 45(10): 91.

[6] Qianzi Shen, Zijian Wang, Yaoru Sun. 2017. Sentiment analysis of movie reviews based on CNN-BLSTM. In Intelligence Science I: Second IFIP TC 12 International Conference, ICIS 2017, Shanghai, China, October 25-28, 2017, Proceedings 2, Springer International Publishing, 164-171.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30.

[8] Stanford NLP. 2024. STANFORDNLP/Imdb Datasets at Hugging Face. Available: https://huggingface.co/datasets/stanfordnlp/imdb

[9] Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, Yan Dai. 2024. Understanding Adam optimizer via online learning of updates: Adam is FTRL in disguise. arXiv preprint arXiv:2402.01567.

[10] Sebastian Bock, Martin Weiß. 2019. A proof of local convergence for the Adam optimizer. In 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 1-8.