**DREAM Challenge 2022**

**Predicting gene expression using millions of random promoter sequences by Zeta**

***Abstract***

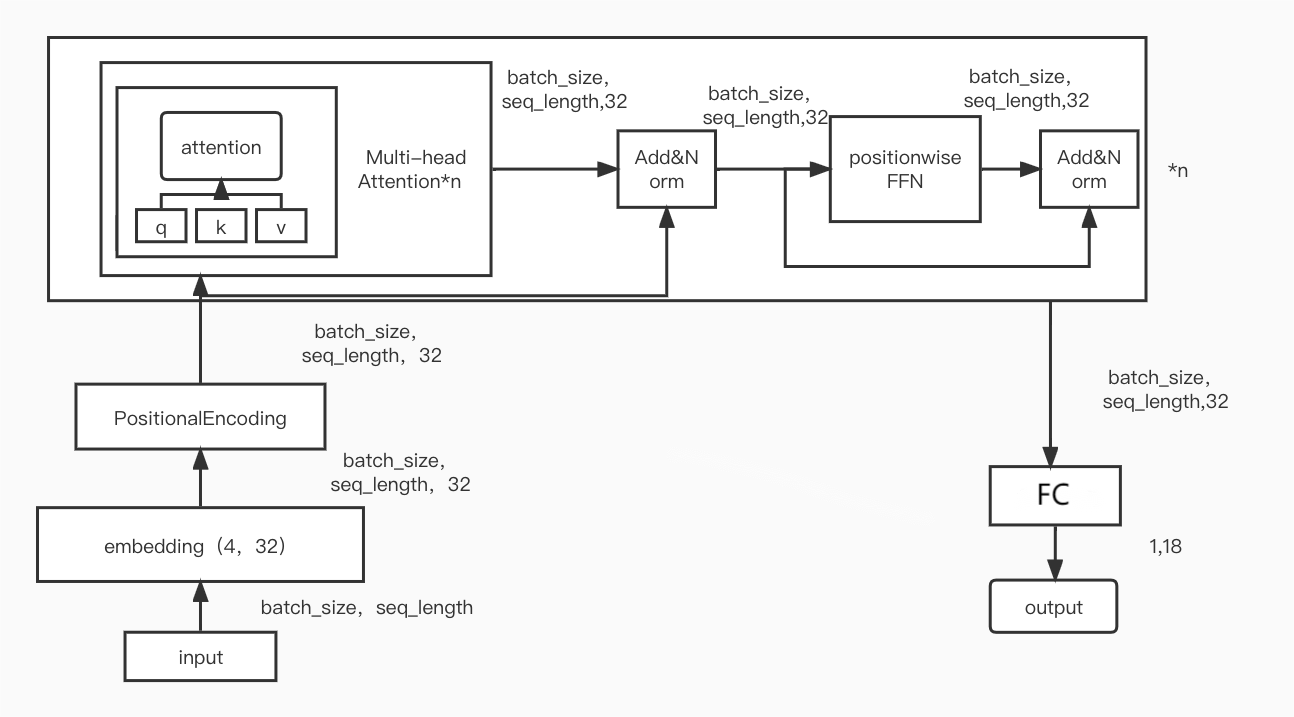
Our model is based on the transformer. In the data embedding part it is divided into two main parts: words embedding and positions encoding. We use the position encoding and encoding parts of the transformer and finally output the probabilities of the eighteen categories through the fully connected layer. We normalise the predictions to eighteen values and then output the category corresponding to the value with the highest probability. Within each category, the probabilities for each sequence are recorded and the final predictions are generated by ranking the probabilities according to the size of the algorithm. For the encoding part, we used a multi-headed attention mechanism, a location-based feedforward network and normalisation.

**1. Description of data usage**

In terms of data processing, we first collated the provided data and filtered out sequences of length 110, which did not contain base N and had integer predicted expression values, to generate a separate file. The collated data were randomly divided into a training set and a validation set in a ratio of 10:1, and training and validation files were generated. Base pairs are represented by numerical codes.

**2. Description of the model**

The number of input sequences is stored via the batch\_size variable, and the data obtained after embedding is (batch\_size, seq\_length,32). After n-layer encoeder with Multi-head Attention inside, the final output is (1, 18). After assigning the sequence to the classification with the highest probability, the final prediction is then obtained by ranking the probabilities by their size. The loss function uses CrossEntropyLoss and the optimiser uses SGD. the fully connected layers involved in the network are all biased.



**3. Training procedure**

After inputting the data, the model calculates the loss of the training set, the loss of the validation set and the F1-Score while calculating and outputting the corresponding predictions, and determines the effectiveness of the training based on these two values. The loss function is CrossEntropyLoss, the optimiser is SGD, theregularisation is 0.9, and the learning rate is 0.03. The F1-Score for the validation set is 0.29.

**4. Other important features**

For positional encoding, we used a form of sine cosine encoding. The Q,K,V matrices connoted in the multi-headed attention mechanism were given the same input as well as different weights. The four bases were expanded to 32 dimensions in Word Embedding.

**5. Contributions and Acknowledgement**

**5.1 Contributions**

|  |  |  |
| --- | --- | --- |
| **Name** | **Affiliation** | **Email** |
| Chen JiaXing | Beijing Normal University-Hong Kong Baptist University United International College | jiaxingchen@uic.edu.cn |
| Wang JingZhe | Beijing Normal University-Hong Kong Baptist University United International College | wang.jingzhe@foxmail.com |
| Yang AiWei | Beijing Normal University-Hong Kong Baptist University United International College | Lorenzoyang18@gmail.com |

**5.2 Acknowledgement**

Thanks to Ms. Chen Jiaxing for her careful guidance and the server resources provided by Beijing Normal University-Hong Kong Baptist University United International College.

**6. References**

[1] Vaswani, A. , et al. "Attention Is All You Need." arXiv arXiv, 2017.

**7. Feedback (optional)**