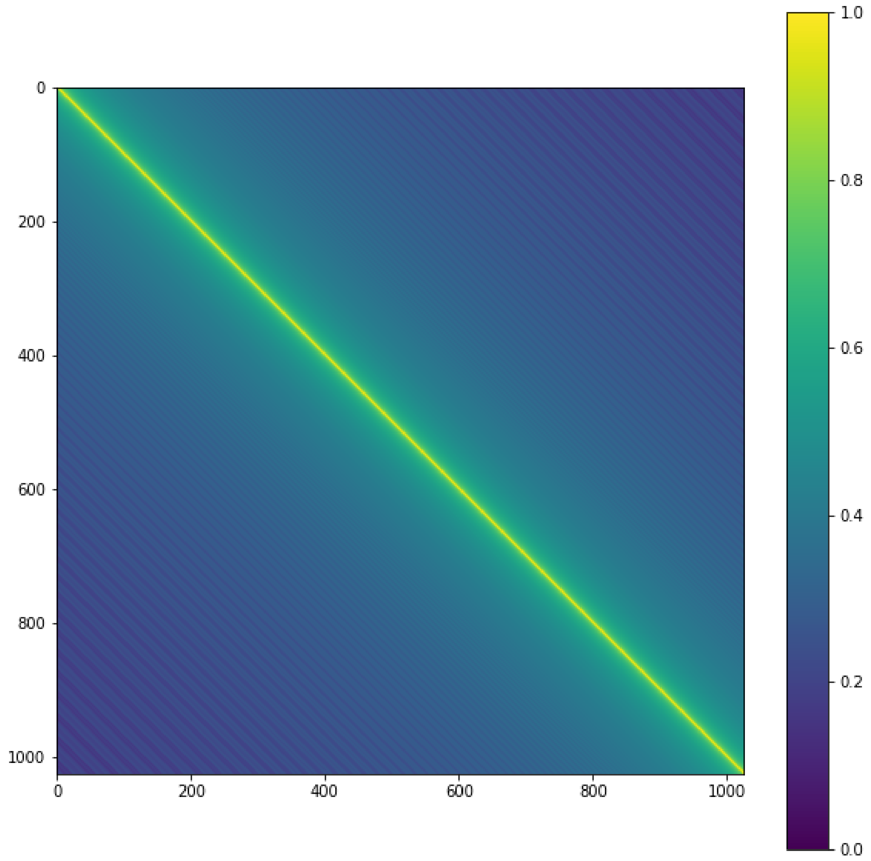
ML HW5 Report

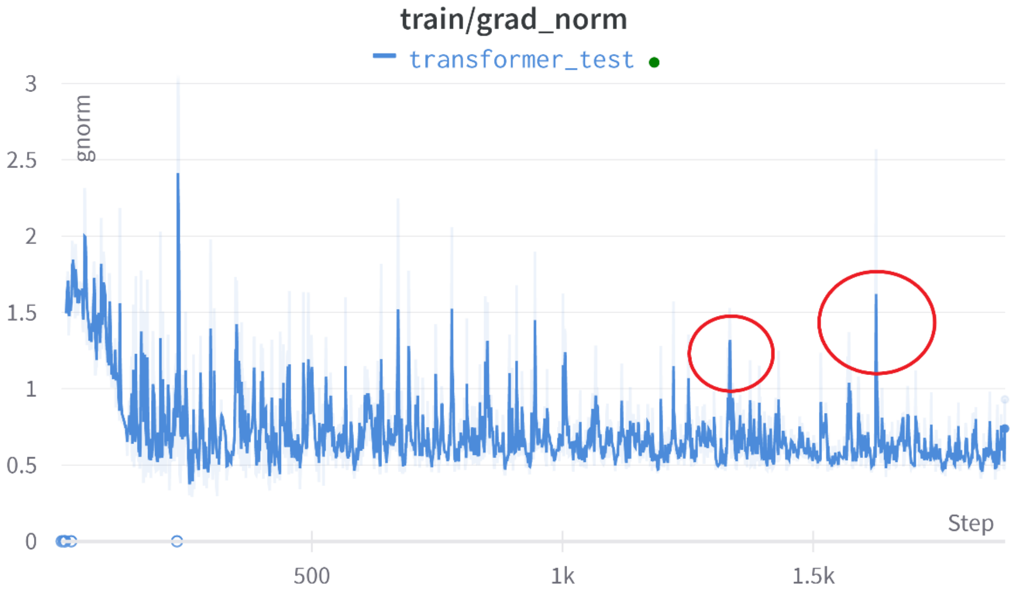
物理三 潘阜承 B08202036

1. Below is the figure of similarity matrix for positional embedding in the Transformer decoder. To be particular, I used Cosine Similarity to calculate the similarity between two vectors.



As we can see from the figure, the similarity values are the highest in the diagonal of the matrix, which is what we expected since they are the similarity of the two identical vectors. Moreover, the figure is highly symmetric, which is also reasonable since the embedding function used in transformer are cosine and sine functions. Last but not least, it is also obvious that the similarity decay periodically toward the bottom-left and top-right corner. This is also due to the characteristic of cosine and sine function.

2. Below is the figure of gradient norm versus the training steps (1st epoch).



As we can see, the gradient is fairly unstable, and there are several places where the norm of the gradient explode (exceeding the clipping value 1).