ML HW4 Report

B08202036 潘阜承 物理三

1. Make a brief introduction about a variant of Transformer.

I would like to introduce “Reformer” whose full content can be refer to the reference below. The motivation of Reformer is that, although Transformer is able to achieve state-of-the-art performance, it is usually computationally expensive, especially on long sequences of data. Thus, Reformer aims to improve the efficiency of the Transformer. First of all, Reformer replace the dot-product attention in the standard Transformer with the one that uses locality-sensitive hashing. Briefly saying, they let query Q and key K be the same since this has little effect on the performance. Then, since the softmax is dominated by the largest term, for each query, they only pick out the keys which are closest to it. This way, they are able to reduce the parameters in Transformer. Second, they use reversible residual layers instead of standard ones, which allows storing activations only once in the training process instead of N (number of layers) times. These two methods enable Reformer to reduce the memory usage as well as the training time while retaining the same performance as the standard Transformer.

2. Briefly explain why adding convolutional layers to Transformer can boost performance.

The main reason why convolutional layers are able to enhance the performance of Transformer lies in the fact that they are able to extract local features effectively while standard Transformer is less capable of capturing local information. Thus, by combining convolutional layers with standard Transformer, we are allowed to extract both local and global features effectively, which in turn leads to enhanced performance.