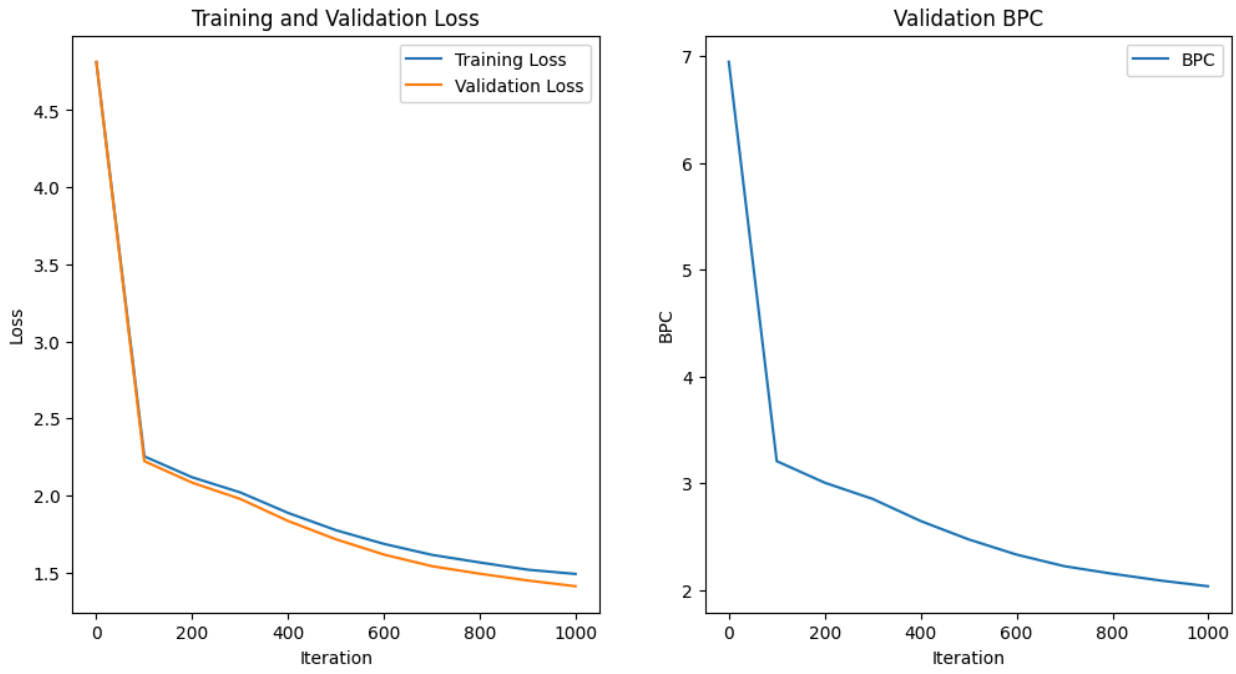
CHKTAP011 - NLP Assignment 2

Implementation

Using minGPT as a reference, I developed a decoder only language model. Each decoder block consists of a multi-head attention block and feed-forward neural networks. Dropout layers were added within the attention and feed-forward blocks.

Fine-tuning

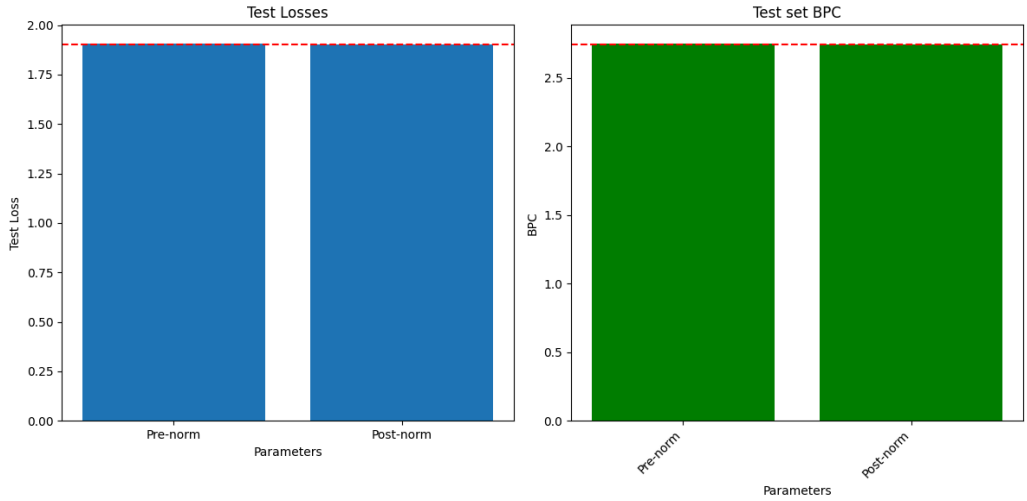
During fine tuning, each model was trained for 1000 epochs. I then compared model performance to determine the optimal values for each parameter tested. The two metrics used to measure performance were bits per character (BPC) on the validation set and the training time.The addition of time as a metric came from the observation that training time had a larger variance than BPC. Favoring parameters that result in shorter training times adds an element of resource efficiency to the model evaluation. An example of the results



Experiments

I added a learning rate scheduler in an attempt to further speed up training. This took the form of a OneCycleLR scheduler. This type of scheduler starts by increasing the learning rate at the start of training but reduces it towards the end to speed up convergence. As a result, the model was able to achieve a BPC of 2.0465 on the validation data after only 1000 epochs of training

I initially trained the model using pre-normalisation on the self-attention and feed-forward blocks. This normalizes input values before feeding them to each block. Doing so is believed to stabilize training in deeper networks. Keeping all other hyperparameters the same, I experimented to see the effect of instead normalizing after each block. The results of this experiment are shown below and revealed identical performance for both normalization methods around 2.7.



Conclusions

Adding the scheduler made the most difference in BPC. Normalisation method seemed to be insignificant and given enough time to train, any model could possibly achieve good performance