HW5

Problem 1 Using the "UCLA is at Westwood" example, explain the GPT model at the word "at" based on the residual stream. Assume there is an embed layer, two layers of attention and MLP, followed by the unembed layer. Please write with concrete vectors and matrices. You only need to use a single head at each attention layer, and no need to worry about position embedding, layer norm, and drop out etc. Please explain the model in terms of retrieval from context and retrieval from memory. Please specify the concrete dimensions of all the vectors and matrices in your design, including the size of the vocabulary, indicate all the parameters to be learned, and calculate the number of parameters in your design. Please also elaborate on parallelizing the back-propagation calculation, although you do not need to provide full mathematical details.

Problem 2 Please play with the PyTorch code provided by the following webpage:

`https://github.com/karpathy/nanoGPT`

Write a brief explanation of the code and show your results. You can explore the code by varying the design parameters.

Problem 3 Please read book chapters 1-5. For each chapter, write a brief review.