Reduce Transformer model's CPU and Memory Loadings 作業

隨著大型語言模型參數量的增加和序列長度的擴展，Transformer 模型的局限性逐漸顯現，尤其是自注意力機制的（Self-Attention Mechanism）的計算量和內存需求隨著序列長度的增加而呈現平方級 (subquadratic) 的增長，成為一個顯著的瓶頸。

本次作業目標在於使用 Mamba 完成對 Transformer 模型的優化，減少模型在訓練和推理過程中對 CPU 和內存的負載。

- 作業目標:
  - 任務一：自行建立一個 Transformer 模型，並將此模型用於預測台積電（2330）的明日股價
    - 訓練集區間：2010/01/01 ～ 2023/12/31
    - 模型推論時的預期 Input/Output：
      Input：該隻股票自 t-29 日到 t+0 日之間每一個開市日的價格狀況。
      Output：模型預測出 t+1 日時該隻股票的價格狀況。
      （價格狀況 = 開盤價、收盤價、最高價、最低價）
  - 任務二：使用 Mamba 架構對 Transformer 模型進行優化：
    Step1. 將剛剛在任務一創建的 Transformer 模型進行優化，修改成 Mamba 架構。
    Step2. 嘗試 tuning Mamba 讓模型的效果更好。
  - 任務三：比較優化過後的模型以及原始模型的大小、準確率、推論效率之差異：
    - 比較兩個模型的大小差異（單位為 MB）
    - 進行回測（backtesting），比較兩個模型推論的準確率：
      回測方式：
      Step1. 選定一個台股有開市的日期當作 t+1 日（e.g. 2024/03/29），並蒐集此日的價格狀況，這將會作為模型推論輸出的正確答案。
      Step2. 蒐集 t-29 日至 t+0 日（e.g. 2024/02/28~2024/03/28）每一個開市日的價格狀況，並作為模型推論時的 Input。
      Step3. 將兩個模型都輸入同樣的 Input，並得到模型推論的結果（Output）。
      Step4. 將兩個模型的 Output 與 Step1. 得到的答案進行比較，並計算出 Output 值與答案的 RMSE（均方根誤差）。
    - 比較兩個模型的推論效率，以 second 為單位。
- 應繳交:
  1. 操作手冊: 這個手冊應詳細說明如何優化模型以及執行推論，請使用 Step By Step 的方式說明。
  2. Docker Container: 請將您的系統建置在 Docker Container 內，並製作一個 start.sh，讓助教直接呼叫 start.sh 就可以開始測試您開發的系統。
  3. Codes: 原始程式碼，以及原始的 Transformer based 模型檔、優化過後的 Mamba 模型檔。
  4. 此作業的 word 檔報告: 報告中至少應附上任務一中創建的 Transformer 模型推論時的截圖、任務二中對模型優化的過程敘述以及截圖，以及任務三中的各項測試數據以及截圖。
- 驗收時間: 2024 年 3 月 31 日
- 評分標準：
  - 任務一 25%

- 任務二 25%
- 任務三 25%
- 作業的 word 檔報告 25%

- Mamba 論文：https://arxiv.org/abs/2312.00752
- Mamba 開源程式碼：https://github.com/state-spaces/mamba