**Lab5 Report**

**1. Testcases**

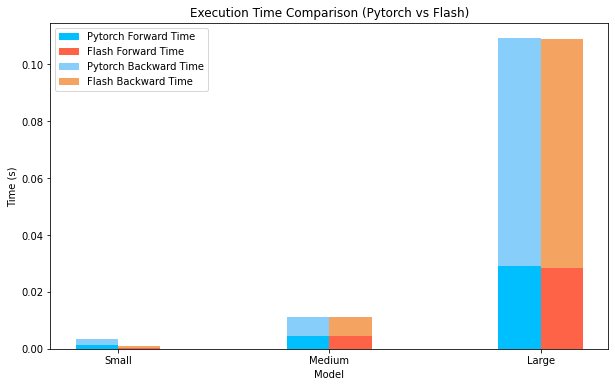
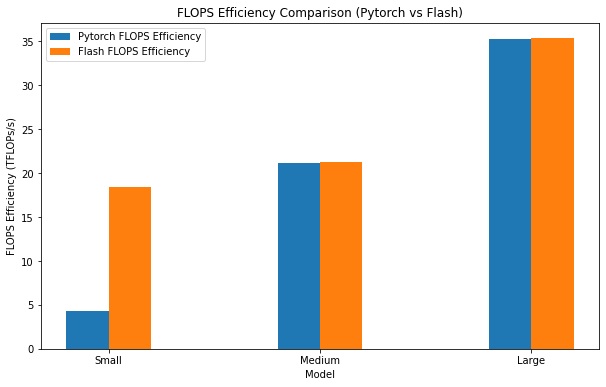
分別對pytorch、FlashAttention2使用small、medium、large三種測資：

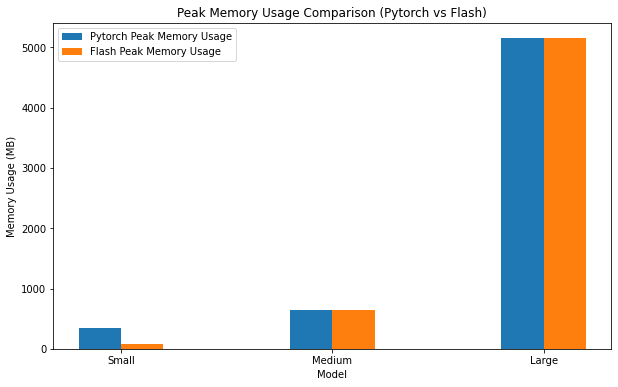
Small: --batch\_size 16 --seq\_len 512 --num\_heads 8 --emb\_dim 512 --causal --repeats 10

Medium: --batch\_size 32 --seq\_len 1024 --num\_heads 16 --emb\_dim 1024 --causal --repeats 20

Large: --batch\_size 64 --seq\_len 2048 --num\_heads 32 --emb\_dim 2048 --causal --repeats 30

**2. Plots**





**3. Discussion**

1. Execution Time：在小模型上，Flash 顯著快於 Pytorch，但在另外兩個的差異不大。
2. FLOPS Efficiency：Flash 在小型模型的 FLOPS 效率遠高於 Pytorch，在中型和大型模型中表現差不多。
3. Peak Memory Usage：Flash 在小型模型的記憶體使用顯著低於 Pytorch，而在中型和大型模型的使用量接近。

整體來說，FlashAttention的效益在小模型上較能顯現。