HW4 – Part 2

Summary of Results:

Experiment	Learning Rate	Batch Size	Dropout Probability	#Blocks	#Heads	Hidden Dim	NDCG@10	HR@10
Base Case	0.001	128	0.5	2	2	50	0.2818	0.4471
Case 1	0.0005	128	0.5	2	2	50	0.2811	0.4438
Case 2	0.001	256	0.3	2	2	50	0.2885	0.4449
Case 3	0.001	128	0.3	3	2	50	0.2852	0.4521
Case 4	0.001	128	0.7	2	2	50	0.2687	0.4314
Case 5	0.001	128	0.3	2	4	50	0.2894	0.4478

Insights and Observations:

1. Effect of Learning Rate (Base Case vs. Case 1):
    a. Learning Rate Change: Reduced from 0.001 to 0.0005.
    b. Observation: Slightly worse performance in NDCG@10 and HR@10.
    c. Interpretation: A lower learning rate can slow down convergence, preventing the model from fully leveraging the data during the training epochs. This suggests 0.001 is a better learning rate for this setup.

2. Effect of Batch Size and Dropout (Base Case vs. Case 2):
    a. Batch Size: Increased from 128 to 256.
    b. Dropout Probability: Reduced from 0.5 to 0.3.
    c. Observation: Improved NDCG@10 but no significant change in HR@10.
    d. Interpretation: A larger batch size can stabilize gradient updates, leading to better generalization, while a lower dropout rate helps the model learn more expressive representations, as seen in the improved ranking metric (NDCG@10).

3. Effect of Number of Blocks (Base Case vs. Case 3):
    a. Number of Blocks: Increased from 2 to 3.
    b. Observation: Higher HR@10 and slightly better NDCG@10.
    c. Interpretation: Adding more blocks enhances the model's capacity to capture sequential patterns, improving hit rate, especially for longer sequences.

4. Effect of Dropout Probability (Base Case vs. Case 4):
    a. Dropout Probability: Increased from 0.5 to 0.7.
    b. Observation: Both metrics decreased significantly.
    c. Interpretation: Higher dropout probability may lead to underfitting, as the model discards too much information during training, reducing its ability to learn from the data.

5. Effect of Attention Heads (Base Case vs. Case 5):
    a. Number of Attention Heads: Increased from 2 to 4.
    b. Observation: Best performance in NDCG@10 and slightly better HR@10 compared to the base case.
    c. Interpretation: Increasing the number of attention heads allows the model to capture finer-grained interactions in the sequence, leading to better rankings of the next items.

Overall Insights:

1. Dropout probability needs careful tuning. While reducing it to 0.3 helped (Case 2, Case 5), increasing it to 0.7 was detrimental (Case 4).

2. Number of blocks and attention heads significantly impact model performance. Adding more blocks and heads can improve metrics, but this likely comes with a computational tradeoff.

3. Batch size plays a role in stabilizing training and improving rankings, as seen in Case 2.

4. Learning rate around 0.001 works well, consistent with the SASRec paper, as too low a value slows convergence without apparent benefits.

Recommendations:

1. Further experiments with combinations of higher attention heads, more blocks, and moderately lower dropout rates may yield better results.
2. Evaluate computational efficiency and training time tradeoffs when scaling parameters, such as the number of heads or blocks.