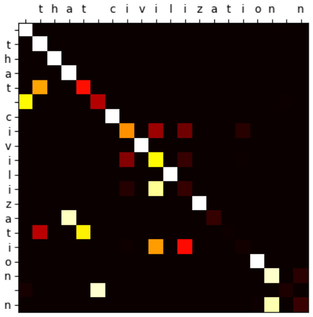
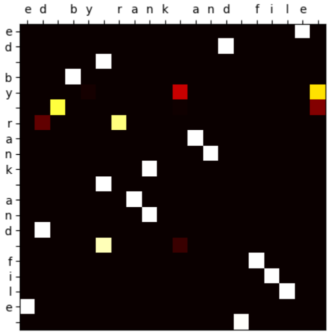
CS6120
Fall 2024
Homework 3
Jiayue Zhang

Part I

Q2



In this attention masks, the model places a strong focus along the diagonal, meaning each position primarily attends to itself and nearby positions. This suggests that the model is focusing on local context, which aligns well with its task of counting character occurrences.



However, in this case, attention distributed across other parts of the sequence, which may indicate the model is learning to relate distant occurrences of the same character. This broader attention behavior is expected as it enables the model to capture repeated characters across the sequence.

Q3

When applying more layers, not all attention masks may exhibit the expected pattern of focusing on prior occurrences of the same character.

When applying “less clear” attention masks, the diffuse attention gets broader, where tokens attend to a wider range of positions, including those that may not be directly relevant to counting occurrences. Since higher layers are capturing more abstract or global patterns across the sequence, rather than focusing solely on individual character repetition, the model is balancing local character-level attention with more general sequence-level understanding.