Mathematical Explanation of Ghost Attention (GhostAttn)

Ghost Attention (GhostAttn) is an advanced technique used to optimize the traditional self-attention mechanism in transformer models, making them more efficient both in terms of computation and memory usage.

1. Traditional Self-Attention Mechanism:

In traditional self-attention, each token interacts with every other token to determine its attention score. This involves computing attention scores for every pair of tokens, which is computationally expensive.

Mathematically:

Attention(Q_i, K_j) = (Q_i . K_j) / sqrt(d_k)

where Q_i, K_j are the query and key vectors, and d_k is the dimension of the key vectors.

2. Ghost Attention Mechanism:

GhostAttn introduces "ghost tokens," which represent groups of tokens. The attention mechanism then operates on these ghost tokens instead of all individual tokens, reducing the computational load.

Mathematically:

Grouping Tokens into Ghost Tokens:

g_i = sum(w_ij . x_j) for x_j in Group i

Applying Self-Attention on Ghost Tokens:

Attention(Q_g_i, K_g_j) = (Q_g_i . K_g_j) / sqrt(d_k)

Final Output:

Output_i = sum(Attention(Q_g_j, K_g_j) . V_g_j) for g_j in G

3. Advantages of GhostAttn:

- Reduced Computational Complexity: From O(n^2) to O(nm + m^2)
- Reduced Memory Usage: Similarly reduced from O(n^2) to O(nm + m^2)
- Scalability: Allows transformers to handle longer sequences or larger datasets efficiently.

In summary, GhostAttn optimizes the attention mechanism by introducing intermediate ghost tokens, significantly improving the efficiency and scalability of transformer models.