

Short Version

Motivation: Quantized Transformer without Dot-product and Softmax.

Basic idea: Replace Dot-prod with Manhattan and Softmax with ReLU

$$QK^T \rightarrow \sum_k |Q_{ik} - K_{jk}|$$
$$\sum_j \text{Softmax}(Z_{ij}) V_{jk} \rightarrow \sum_j (V_{jk} - Z_{ij}^+)^+$$

Observations:

- Reminiscent of subtractive inhibition in biological neurons.
- Removes variable multiplication and require only *half* precision.

Result:

- Comparable training capacity to the conventional mechanism.
- Reduced precision requirements translate into computational efficiency.
- Substantial gains under FHE by avoiding ciphertext multiplication.

Potential:

- Natural integer quantization for deployment under resource constraints.
- May enable end-to-end encrypted Transformer applications.

Future work: Train larger models like BERT, GPT and Vision Transformer.

