In Short

Motivation: Quantized Transformer without Dot-product and Softmax.

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

$$QK^T \to \sum\nolimits_k |Q_{ik} - K_{jk}|$$

$$\sum\nolimits_j \mathrm{Softmax}\left(Z_{ij}\right) V_{jk} \to \sum_j \left(V_{jk} - Z_{ij}^{+}\right)^{+}$$

Observations:

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

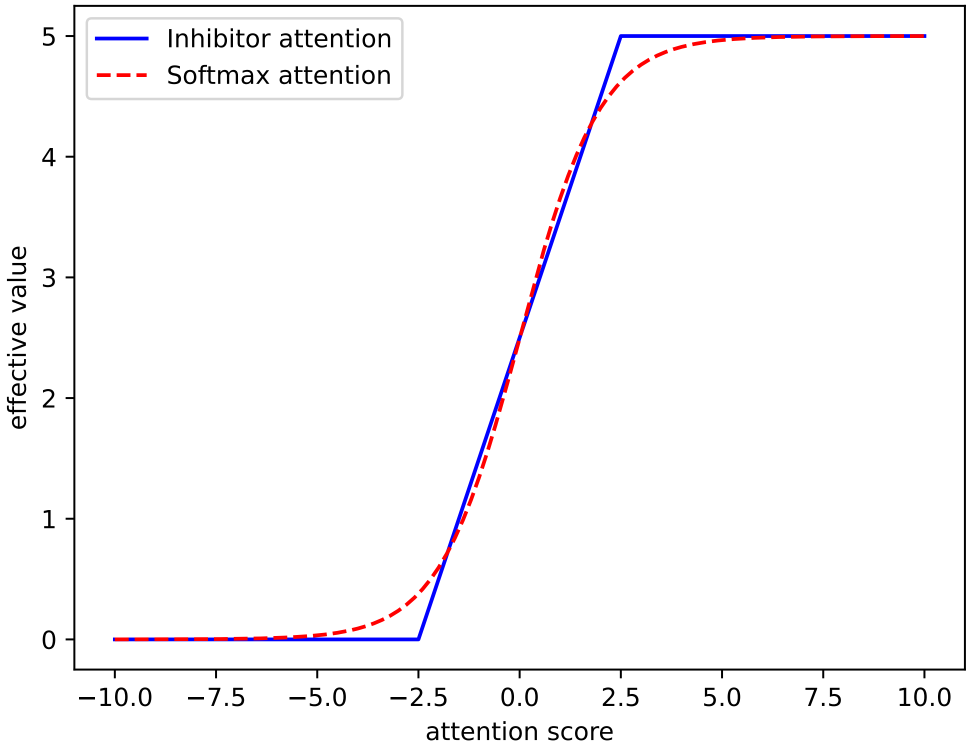
Result:

- Comparable training capacity to the convetional mechanism.
- Reduced precision requirements translate into computational efficiency.
- Substantional 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.


Inhibitor attention
Softmax attention
effective value
attention score