In Short

Basic idea: Let $u_t = Wx_t + Uh_{t-1} + b$ and change gated RNNs according to

$$\sigma(u_t) \odot h_{t-1} \rightarrow \left(u_t^- + h_{t-1}\right)^+$$

such that multiplication is replaced with addition and sigmoid with ReLU.

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 MLM and Vision Transformer.

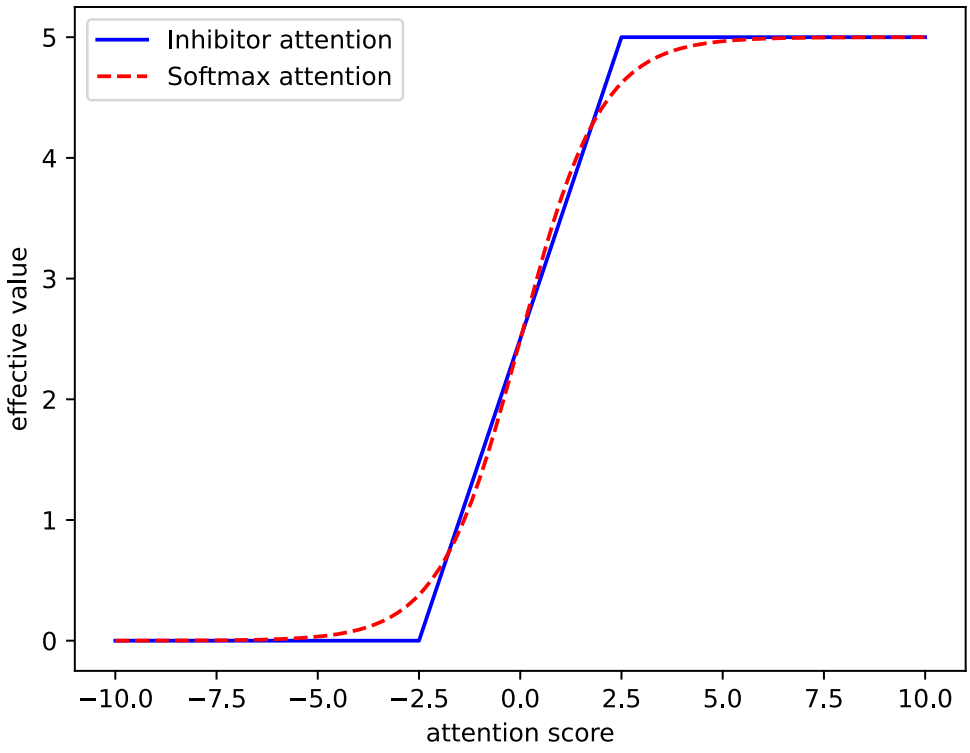

Figure 1: Comparison of the Inhibitor mechanism with conventional attention.