The Inhibitor: ReLU and Addition-Based Attention for Efficient Transformers under TFHE

Rickard Brännvall[1] and Andrei Stoian[2]

[1]Computer Science Department, RISE Research Institutes of Sweden, rickard.brannvall@ri.se
[2]Machine Learning Group, Zama, Paris, France, andrei.stoian@zama.ai

(poster) (arxiv)

To enhance the computational efficiency of quantized Transformers, we replace the dot-product and Softmax-based attention with an alternative mechanism involving addition and ReLU activation only. This side-steps the expansion to double precision often required by matrix multiplication and avoids costly Softmax evaluations but maintains much of the core functionality of conventional dot-product attention. It can enable more efficient execution and support larger quantized Transformer models on resource-constrained hardware or alternative arithmetic systems like homomorphic encryption.

In Short

Motivation:

- Quantized Transformer that doesn't use Dot-product and Softmax.

Basic idea:

- Replace Dot-prod with Manhattan distance and Softmax with ReLU

$$QK^T \rightarrow \sum_k |Q_{ik} - K_{jk}|$$

$$\sum_j \text{Softmax}\left(Z_{ij}\right) V_{jk} \rightarrow \sum_j \left(V_{jk} - Z_{ij}^{+}\right)^{+}$$

Observations:

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

Result:

- Comparable training capacity to the conventional 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.

Related work:

- Inhibitor gate is efficient also for GRU and LSTM.

Future work:

- Train larger models like BERT, GPT and Vision Transformer.

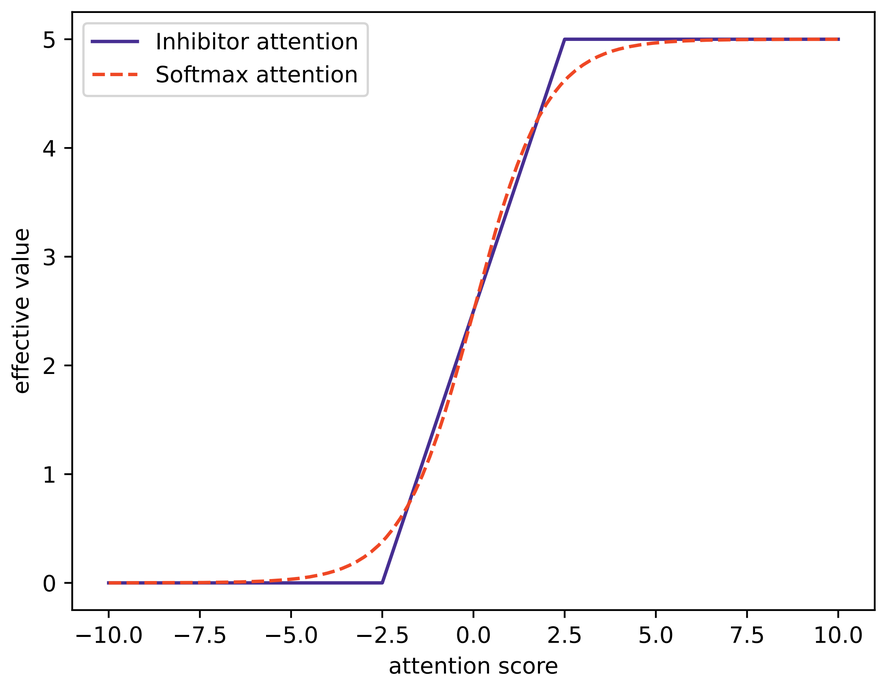

Figure 1: Comparison of the Inhibitor gate with the Sigmoid function and multiplication based gate used for the conventional GRU and LSTM.

Training Results

Training experiments on four common benchmark tasks show test set prediction scores comparable to those of conventional Transformers with dot-product attention.

Benchmark Tasks	Adding	MNIST	IMDB	IAMW
Dot-prod Attention	0.11%	98.2%	87.2%	17.9
Inhibitor Attention	0.12%	97.9%	87.3%	18.1

Table 1: The Inhibitor shows comparable performance to conventional attention for Transformers trained on four standard tasks (for mse, acc, acc, and edit distance, respectively).

Scaling Experiment

Our scaling experiments also suggest significant computational savings, both in plaintext and under encryption.

Timing Plaintext	32	64	128	256
Dot-prod Attention	98.6 µs	330 µs	1.2 ms	4.48 ms
Inhibitor Attention	63.1 µs	178 µs	577 µs	2.5 ms

Table 2: Estimated plaintext execution time on CPU for four different sequence lengths (with fixed size single head).

Timing Encrypted	2	4	8	16
Dot-prod Attention	2.68 s	22.4 s	107 s	828 s
Inhibitor Attention	0.749 s	8.56 s	23.8 s	127 s

Table 3: Estimated encrypted execution time on CPU for four different sequence lengths (with fixed size single head).

Circuit Size

The Inhibitor mechanism requires fewer bits precision and avoids expanding the circuit size compared to the conventional dot-product attention mechanism.

TFHE Parameters	T	lweDim	baseLog	level	polySize	int	uint
Inhibitor Attention	2	816	23	1	2048	5	4
Dot-prod Attention	2	817	23	1	2048	6	7
Inhibitor Attention	4	875	22	1	4096	6	5
Dot-prod Attention	4	834	23	1	2048	7	7
Inhibitor Attention	8	795	22	1	4096	5	5
Dot-prod Attention	8	792	22	1	4096	7	8
Inhibitor Attention	16	883	22	1	4096	6	6
Dot-prod Attention	16	794	15	2	4096	8	8

Table 4: Parameters and circuit bit size set by Concrete TFHE compiler for four different sequence lengths (T) processed by the two alternative Transformer attention mechanisms. Note the difference in bit-precision that is required (last two columns).

Conclusion

The ReLU and addition-based attention mechanism examined in this paper may enable privacy-preserving AI applications operating under homomorphic encryption by avoiding the costly multiplication of encrypted variables.

RISE Research Institutes of Sweden
www.ri.se / info@ri.se

CONTACT
Rickard Brännvall, Department of Computer Science, Digital Systems division, RISE, rickard.brannvall@ri.se
RISE is an independent state-owned research instiute with over 3000 employes supporting research and innovation in Sweden.