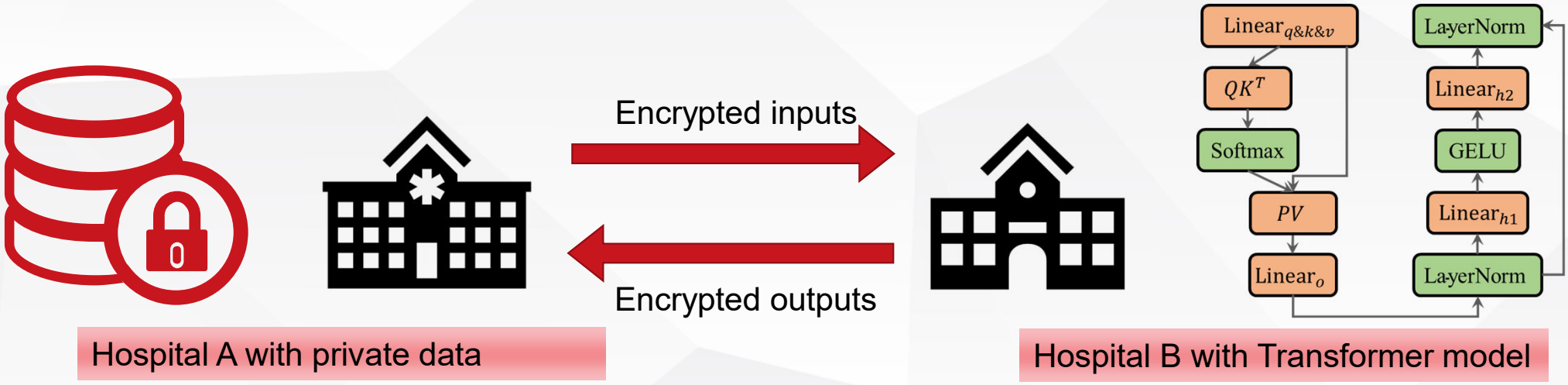
Nimbus: Secure and Efficient Two-Party Inference for Transformers

Zhengyi Li, Kang Yang, Jin Tan, Wen-jie Lu, Haoqi Wu, Xiao Wang, Yu Yu, Derun Zhao, Yancheng Zheng, Minyi Guo, Jingwen Leng

hobbit@sjtu.edu.cn

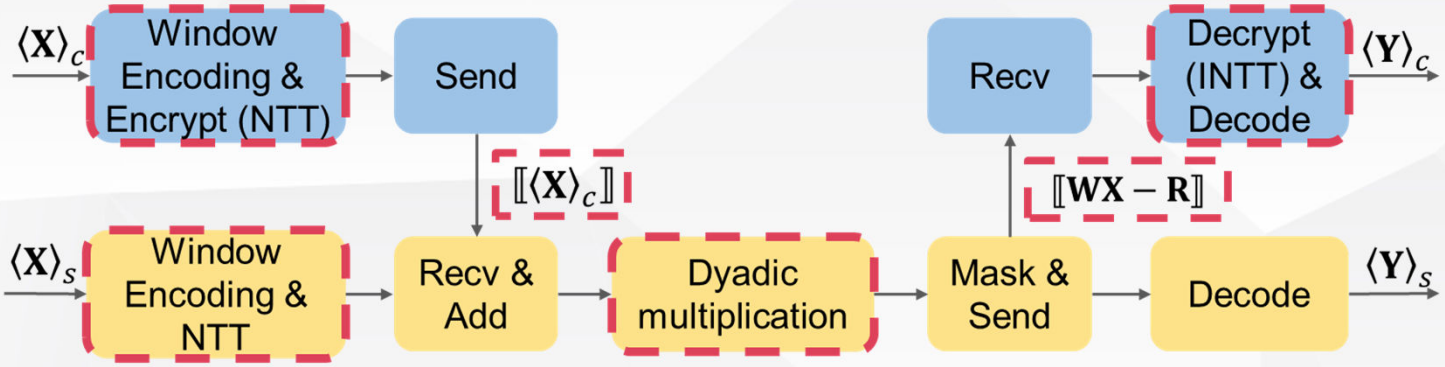
2024.11

Secure Inference of the Transformer Models



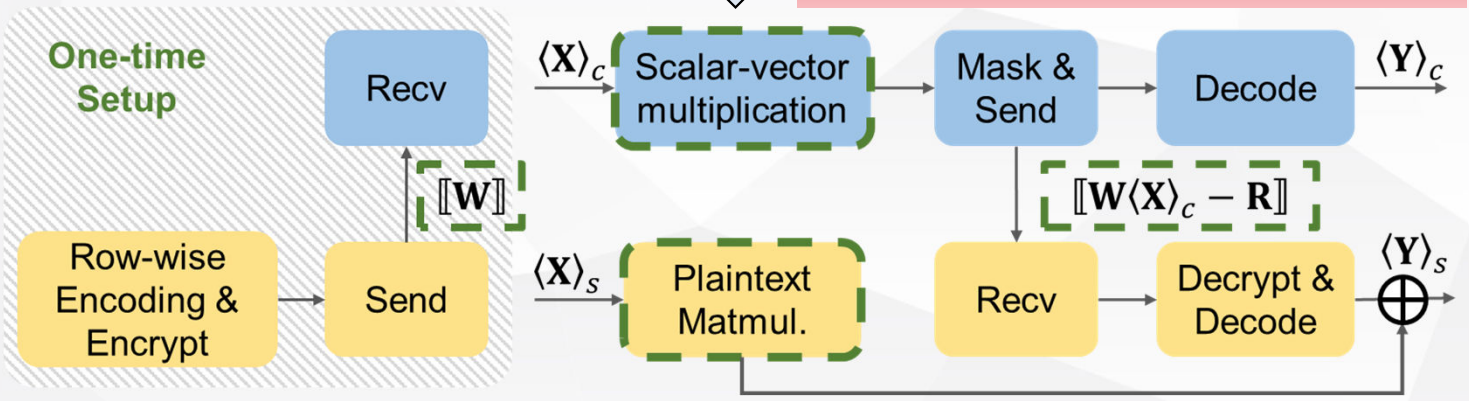
- On varying tasks, Transformer models have demonstrated promising performance.
  - ChatGPT, medical application
- However, the secure inference of the Transformer model faces efficiency challenge.
- Nimbus focuses on the two-party secure inference.

Nimbus—Linear Protocol



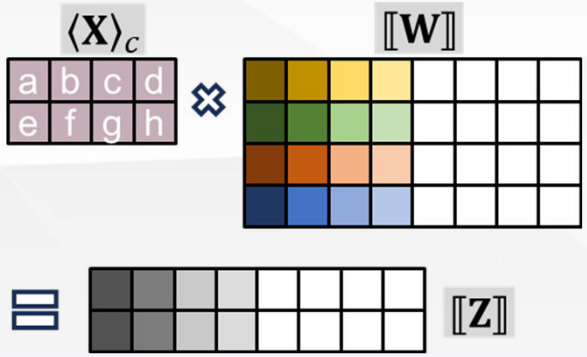
Server-side inner product protocol

Client-side outer product protocol



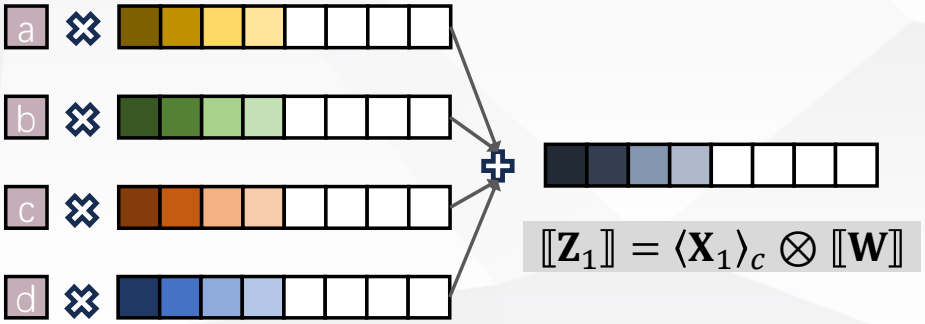
- Traditional intuition: client sent encrypted data to the server for computation
- Utilizing the static nature of model weights.
- Save the input communication of the input ciphertext.
- Keep similar computation workload of the client.

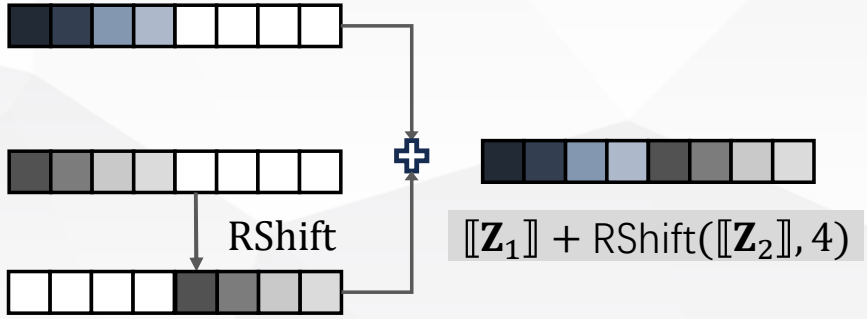
Nimbus—Linear Protocol



- Without the restriction of input communication allows a new encoding strategy
  - Achieve efficient multiplication of scalar and ciphertext .
  - Compact output ciphertext through “free” right-shift packing.

Use activation sharing as scalar directly

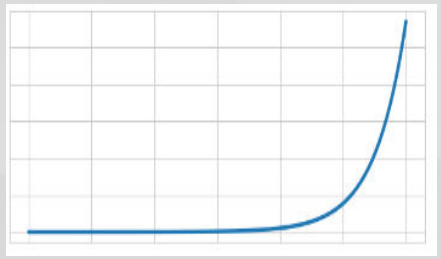




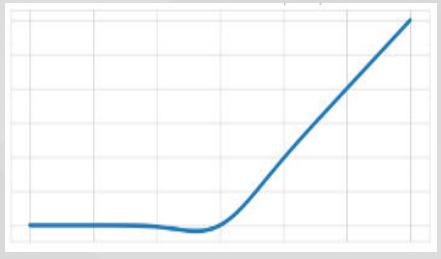
Encoding weights into ciphertext in a row-wise fashion

Achieve compact output ciphertext

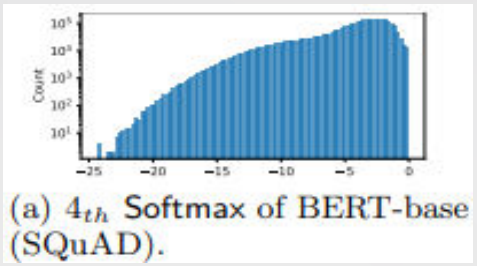
Nimbus—Approximation of Nonlinear Functions



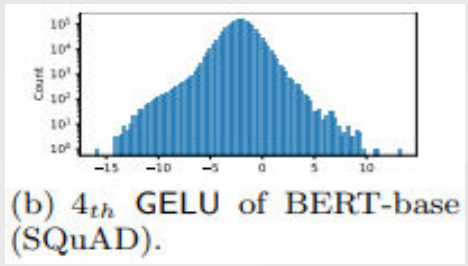
Exponential function



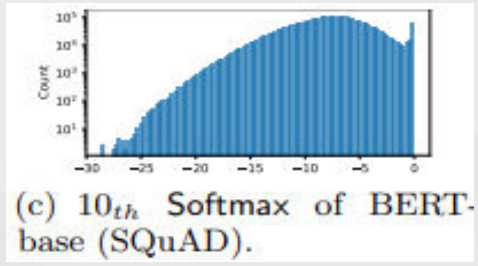
GELU function



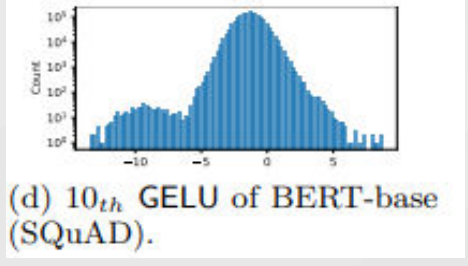
(a) 4_{th} Softmax of BERT-base (SQuAD).



(b) 4_{th} GELU of BERT-base (SQuAD).



(c) 10_{th} Softmax of BERT-base (SQuAD).

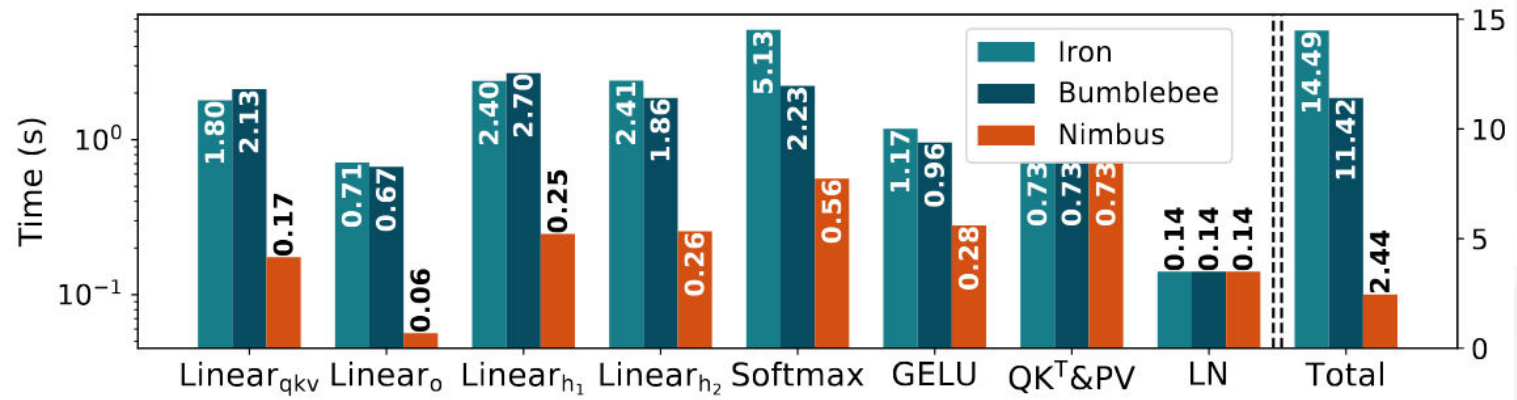


(d) 10_{th} GELU of BERT-base (SQuAD).

- Approximate nonlinear functions through piecewise polynomial for secure evaluation.
- Directly fitting regards inputs as uniform distribution.
- Input distributions of nonlinear functions present regular patters:
  - Softmax: 80% input values of falls in [-5,0];
  - GELU: 90% input values smaller than 0.
- We propose distribution-aware fitting of the piecewise polynomial.
- Achieve low-degree and less-piece piecewise polynomials.

Nimbus--Experiments

- Around 5X overall speedup on BERT-base model:
  - ~10X for linear;
  - ~4X for nonlinear.



- Nimbus has negligible impact on the accuracy.

Method	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	Avg.
	Matthews corr.	Acc.	F1	Pearson	Acc.	Acc.	Acc.	Acc.	
FP baseline	58.63	92.88	90.12	88.24	91.22	84.74	91.28	67.87	83.12
Bumblebee	58.40	92.88	90.12	88.28	91.21	84.74	91.39	67.87	83.11
Nimbus	58.40	92.78	90.42	88.12	90.98	84.37	91.37	67.87	83.04


Thanks!
饮 水 思 源 爱 国 荣 校