A FAST CONVOLUTION ALGORITHM FOR ACCELERATING ENCRYPTED DOMAIN MODEL INFERENCE

Huan-Chih Wang, Ja-Ling Wu

National Taiwan University, {whcjimmy, wjl}@cmlab.csie.ntu.edu.tw

FHE.org Conference 2023

Motivations

- The convolutional layer is the costliest in both plaintext and ciphertext domain inferences.
- The multiplication is much more costly than the addition in the encrypted domain, so the multiplication-intensive functions are even more challenging for private inference.

Contributions

- We apply the fast convolution algorithm (FCA) [2] into the encrypted domain and use the Kronecker product to make FCA more multiplication-efficient than the direct convolution algorithm (DCA).
- As illustrated examples, we test the algorithms with the MNIST and the CIFAR-10 datasets, showing that FCA can save at least 28% inference time than DCA.

Methodology

Fast Convolution Algorithm (FCA)

- When we use the $r \times r$-tap filter to generate $o \times o$ outputs, written as $F(o \times o, r \times r)$, we only require $(o+r-1) \times (o+r-1)$ multiplications at a minimum.
- FCA would be $Y = A^T((GKG^T) \odot (B^T[H]B))A$ in the matrix form, and it needs 4 consecutive multiplications to get the result Y.
- With the Kronecker product, the consecutive multiplication reduces to 2. The matrix form would become $Y = A^T \otimes A(vec(GKG^T) \odot (B^T \otimes Bvec([H])))$.

Protocol 1 The FCA Information Processing Workflow.

Inputs. Alice has a key pair (pk and sk) of an FHE cryptosystem and data D. Bob has the kernel (either K or $[K]$).

Output. Alice gets the plaintext convolutional result Y convolved by the targeting data and the kernel.

The protocol:

1. Alice encrypts data D and sends the encrypted data $[D]$ to Bob.
2. Bob generates FCA's parameter matrices A, B, and G by the parameters o and r.
3. Bob separates D into $\lceil D^w \times D^h/(r-1)^2 \rceil$ tiles, and each pile has $r-1$ elements overlapping with neighboring tiles.
4. Set $\alpha = o+r-1$ and $\beta = o-1$.
5. For each pile $P = [D]_{m:m+\alpha,n:n+\alpha}$ where $1 \le i \le \lceil D^w/r-1 \rceil$, $1 \le j \le \lceil D^h/r-1 \rceil$, $m = (r-1)i$, and $n = (r-1)j$, Bob runs FCA depending on the type of kernel.
   (a) If K, $[Y_{i:i+\beta,j:j+\beta}] = FCA([P], K)$.
   (b) If $[K]$, $[Y_{i:i+\beta,j:j+\beta}] = FCA([P], [K])$.
6. Bob returns the convolved result $[Y]$ to Alice.
7. Alice uses her secret key sk to decrypt $[Y]$, receiving the final plaintext result Y.

Data Embedding Technique

- We follow CryptoNets [1] to embed data into multiple ciphertexts (Figure 2).
- This embedding way can process a significant number of images simultaneously, having a very short amortized time for each plaintext image.

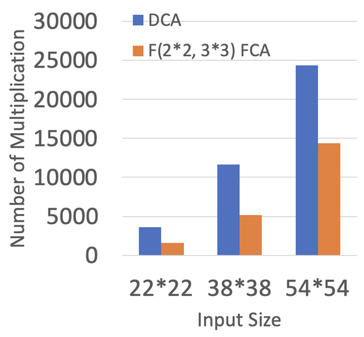

Figure 1: The Number of Multiplication Needed for DCA and $F(2 \times 2, 3 \times 3)$ FCA with a 3×3 Kernel.

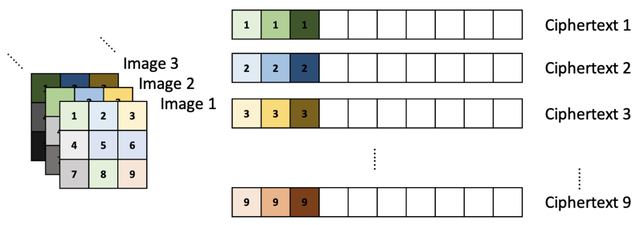

Figure 2: The CryptoNets Embedding Data Technique.

Experiments

Setup

- In all experiments, the FCA parameters are $o = 2$ and $r = 2$.
- We conduct experiments with the MNIST and the CIFAR-10 dataset, respectively.
- The CN model is used for the MNIST dataset, and the CTN model is for the CIFAR-10 dataset.
- We evaluate the inference time of the two convolutional algorithms with the above two models.
- DCA-CN means the CN model with the DCA convolutional algorithm, FCA-CTN is the CTN model with the FCA algorithm, and so on.
- The input data are encrypted in all experiments. However, we also benchmark the inference performance with plaintext and ciphertext model parameters, respectively.

Results

- With plaintext model parameters, FCA saves 28% and 55% overall with the MNIST and the CIFAR-10 datasets, respectively.
- Regarding the ciphertext model parameters, the numbers rise to 40% and nearly 60%.

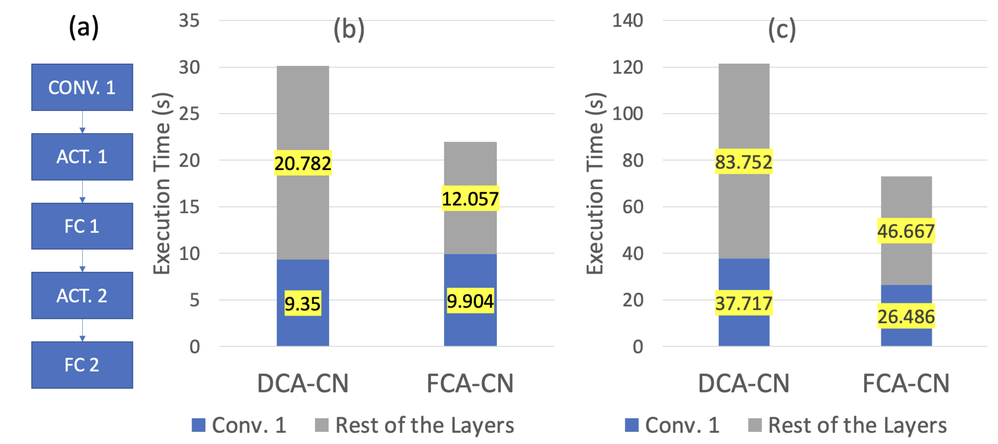

Figure 3: The CN Model Structure and the Inference Time of Different Algorithms with the MNIST Dataset. (**a**) The CN Model Structure, (**b**) Algorithms with the Plaintext Model Parameters, and (**c**) Algorithms with the Ciphertext Model Parameters.

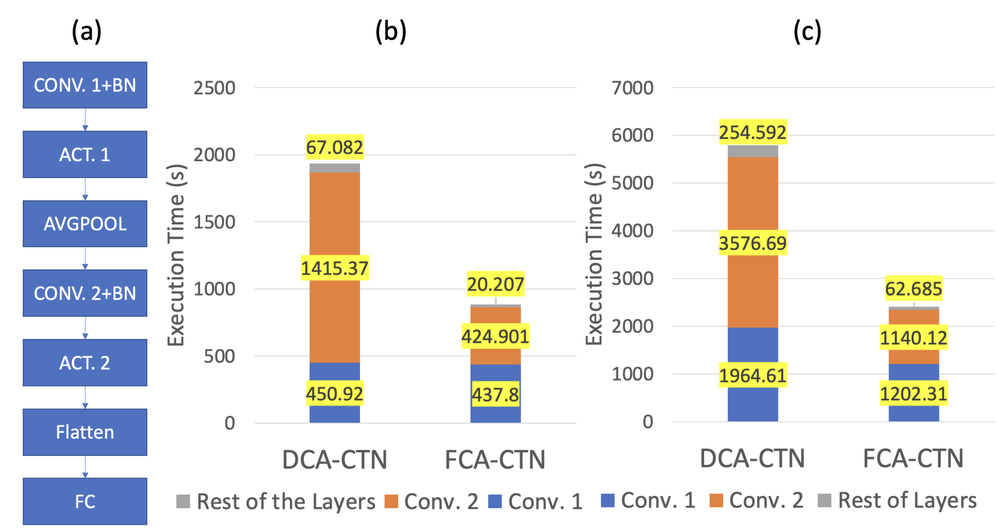

Figure 4: The CTN Model Structure and the Inference Time of Different Algorithms with the CIFAR-10 Dataset. (**a**)The CTN Model Structure, (**b**) Algorithms with the Plaintext Model Parameters, and (**c**) Algorithms with the Ciphertext Model Parameters.

References

[1] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 201–210, New York, New York, USA, 20–22 Jun 2016. PMLR.

[2] A. Lavin and S. Gray. Fast Algorithms for Convolutional Neural Networks. *arXiv:1509.09308 [cs]*, 2015.