동형암호 CNN 가속기의 FPGA 최적화 디자인 설계 연구

이윤지 [1], 이용석 [1], 백윤흥 [1]
[1] 서울대학교 전기정보공학부, 서울대학교 반도체 공동연구소
yjlee@sor.snu.ac.kr, yslee@sor.snu.ac.kr, ypaek@snu.ac.kr

Optimizing FPGA-based CNN Accelerators over Homomorphically Encrypted Data: Based on Survey

Yunji Lee[1], Yongseok Lee[1], Yunheung Paek[1]
[1]Dept. of Electrical and Computer Engineering (ECE) and Inter-University Semiconductor Research Center (ISRC), Seoul National University

요 약

동형암호는 암호화된 상태에서도 연산을 수행할 수 있는 암호이다. 양자컴퓨터 시대가 눈앞에 도래한 지금 양자컴퓨터에도 안전한 격자 문제에 기반을 둔 동형암호가 적용될 수 있는 분야는 무궁무진하다. 특히 최근 AI 의 발전에 따라 프라이버시 문제가 더욱 쟁점이 되고 있는데, 이를 해결하기 위해 등장한 프라이버시 보존 머신 러닝 기술 중 대표적인 것이 동형암호를 이용해 CNN(Convolution Nueral Network)의 추론을 구현하는 것이다. 이 논문에서는 CryptoNets 를 시작으로 HE-CNN 기술의 발전 흐름에 대해 소개하고, FPGA 등 하드웨어 가속 기술을 바탕으로 기존에 적용된 설계 프로세스를 개선한 새로운 디자인 프로세스를 제안한다. 특히 동형암호의 level 변화(연산 횟수 판단)를 고려한 유연성 있는 설계를 위해 다양한 병렬성 파라미터를 도입하고, DSE 로 자원을 최적 분배함으로써 기존보다 더 빠르고 자원 효율적인 가속기를 구현하는 방법을 탐구한다.

1. 서론

암호(Cryptography)란 송신자와 수신자 간 통신과정에서 인가되지 않은 제 3 자가 대화의 내용을 알아내지 못하도록 수학적 기법을 사용하여 원문을 변형시키는 기법이다. 이 원리를 이용하는 DES, AES 등의 기존 암호체계가 현재까지도 사용되고 있으나, 최근 양자컴퓨터의 개발이 현실화되면서 더 이상 AES, RSA 등의 암호체계만으로는 충분히 빠른 것과는 별개로 양자컴퓨터에 대한 안전을 보장할 수 없게 되었다. 따라서 양자컴퓨터에 대해 내성이 있다고 알려진 격자 문제(RLWE)에 기반을 둔 암호를 표준화하고 성능을 개선하기 위한 연구들이 이루어졌다.

오늘 말하고자 하는 대표적인 차세대 암호가 바로 '동형암호(Homomorphic Encryption, HE)' 라는 암호체계이다. '동형(Homomorphic)' 이란 이름은 대수학의 '준동형사상(Homomorphism)' 에서 따온 용어로 두 구조(암호학에선 평문과 암호문 공간)사이의 모든 연산 및 관계를 보존한다는 의미이다.

놀랍게도 동형암호 개념에 대한 첫 제안은 1978 년으로 거슬러 올라간다. 하지만 당시에는 평문이 아닌 암호문의 연산도 동형성을 가졌으면 좋겠다는 제안에 그쳤다면, 2009 년 Gentry[1]에 의해 발표된 최초의 동형암호 알고리즘은 암호문의 동형 연산도 충분히 안전(secure)할 수 있다는 것을 보여주었다. 또한 Gentry 는 재부팅(bootstrapping)이라는 기법을 통해 암호문 연산을 무한히 할 수 있게 하는 완전동형암호(Fully Homomorphic Encryption, FHE)를 제안하기도 했다. 동형암호의 종류에는 부분동형암호(Partial Homomorphic Encryption, PHE), 제한동형암호(Somewhat Homomorphic Encryption, SHE), 그리고 완전동형암호(FHE)가 있는데 PHE 가 덧셈, 곱셈 연산 중 하나만을 지원한다면 SHE 는 덧셈, 곱셈 모두를 지원하나 제한된 수의 연산만 가능하고, FHE 여야만 재부팅을 통해 연산의 수를 무제한으로 확장할 수 있어 최근 개발되는 대부분의 라이브러리는 FHE 를 지원하는 것을 목표로 한다. 대표적인 FHE 스킴에는 각각의 지원 데이터와 연산에 따라 BFV(int), TFHE(bit) 그리고 CKKS(double)가 있다.

동형암호는 우리가 일반적으로 알고 있던 기존의 암호체계와 가장 결정적인 차이가 있다. 그것은 바로 암호문 상태로 연산을 수행한다는 것인데, 이는 아주 명백한 장단점을 가지고 있다. 장점은 물론 우리의 비밀키를 상대방과 공유할 필요가 없어 프라이버시

측면에서는 수요자의 요구를 완벽히 충족할 수 있다는 점이나, 암호문의 연산이라는 평문의 연산과는 비교할 수 없는 연산량의 증가를 보여준다는 치명적인 단점도 가지고 있다. 즉 동형암호(HE), 특히 완전동형암호(FHE)는 아직 상용화하기에는 기존의 암호체계에 비해 매우 느리다는 약점을 해결하기 위해 효율성 개선 연구가 다각도로(알고리즘 최적화, 하드웨어 가속 등) 활발히 진행되고 있다.

따라서 본 논문에서는 먼저 동형암호 기술이 프라이버시 보존 데이터 분석, 대표적으로 기계학습(PPML)에 어떻게 활용되고 있는지 소개하고, 구현 가능한 HE-CNN 아키텍처 설계 및 최적화 기술의 연구 동향을 분석해서 최적화된 하드웨어 기반 설계 자동화 프레임워크를 제안하고자 한다.

2. 프라이버시 보존 기계학습(PPML)에서 동형암호

AI 의 발전이 초 가속화되면서 각종 모델에 제공되는 데이터에 대한 '보안' 이슈가 크게 증가하였다. 우리나라에서는 개인정보로 대표되는 민감데이터가 바로 그것이다. 이런 민감데이터를 보호하기 위한 프라이버시 보존 기술(Privacy Enhancing/Preserving Techniques)에는 동형암호 뿐 아니라 비식별화(De-identification), 차분 프라이버시(Differential Privacy, DP), MPC(Secure Multi-Party Computing), 그리고 연합학습(Federated Learning)등이 대표적으로 알려져 있다.

그 중 동형암호 체계인 HEAAN[2], 현재는 CKKS 라고 알려진 근사계산 동형암호 라이브러리를 이용하면 프라이버시 보존 기계학습(Privacy Preserving Machine Learning, PPML) 등 근사연산을 사용해 데이터 분석을 수행하는 분야에 동형암호의 적용을 통한 강력한 데이터 보호가 가능해졌고, 느린 성능을 보다 최적화한다면 상업성을 기대할 수 있다는 관점에서 연구가 활발히 진행되고 있다. 그리고 HE-CNN 기반의 최적화 핵심인 HE 연산 모듈과 자원 최적화 기법(DSE, HLS 등)은 연산 순서, 병렬 구조, 버퍼 구조 등에 따라 모듈 재배치 및 로직 재설계로 기계학습 외에도 다른 HE 기반 응용 기술(HE-logistic regression, HE-SVM 등)로 확장도 가능하다.

3. HE-CNN 최적화 기술 발전 흐름

이 논문에서는 특히 Convolution Neural Network 에 동형암호 기술이 어떻게 적용되었고, 이를 통해 추론(inference) 과정을 최적화하는 기술의 흐름을 분석하였다. 그리고 기존 연구들을 바탕으로 추가적으로 어떤 기여를 할 수 있을지에 대한 탐구를 진행하였다.

[표 1] HE-CNN 최적화 기술 발전 흐름

	대표 논문	특징
[3]	CryptoNets (2016)	최초의 HE 를 이용한 CNN inference 구현, SIMD packing, square activation ft 사용
[4]	LoLa (2017)	HE 연산을 줄이기 위한 data packing & representation 최적화로 inference 속도 개선
[5]	HEAX (2020)	FPGA 에서 HE 연산을 위한 모듈 단위 분해 및 파이프라인 설계, NTT, INTT, Barrett Reduction 등을 통한 HE 연산 가속, HLS 기반 모듈화
[6]	CoxHE (2022)	
[7]	FxHENN (2023)	RNS-CKKS 기반 CNN 구조에 최적화된 HE 연산 수행, NTT 기반 모듈 병렬화 및 DSE&HLS 로 자동화 설계한 FPGA 프레임워크

위 표는 동형암호를 적용한 CNN 추론 기술의 발전 흐름을 표로 정리한 것이다. 먼저 CryptoNets 가 최초로 동형암호(BFV)를 CNN 추론에 사용하는 데 성공하였다. CNN 의 모든 픽셀을 개별 암호문으로 처리하고 활성함수를 다항함수인 square 함수를 사용하여 HE-CNN 의 가능성을 보여주었다. 하지만 MNIST 기준 205 초가 소요되는 느린 성능으로 추가 최적화가 필요했다. 따라서 이후 데이터 표현 방식 및 packing 방법을 개선한 LoLa 가 등장하였다. 이 논문에서는 dense / stacked / interleaved 및 convolution 표현 방식을 정의하여 dot-product, rotate 연산 등을 최적화하는 방식으로 지연 시간(latency)을 MNIST 기준 2.2 초까지 단축하였다. 이후 FPGA, GPU 등 다양한 하드웨어를 중심으로 병렬화 기반 HE 가속 프레임워크가 제안되었다.[5][6][7][8][9]

[표 2] HE-CNN 추론 성능 비교

모델	플랫폼	Latency (MNIST)	Latency (CIFAR)	특징
CryptoNets	CPU (Xeon)	205s	-	초기 버전, 느림
LoLa	Azure 8vCPU	2.2s	730s	입력/가중치 packing 최적화
Falcon [8]	Azure 8vCPU	1.2s	107s	주파수 도메인 연산
A*FV [9]	GPU 4×V100	5.2s	553.89s	GPU 병렬성 사용
FxHENN (ACU9EG)	FPGA (Low-end)	0.24s	254s	LoLa 보다 9.17× / 2.87× 빠름
FxHENN (ACU15EG)	FPGA (High-end)	0.19s	54.1s	LoLa 보다 11.58× / 13.49× 빠름

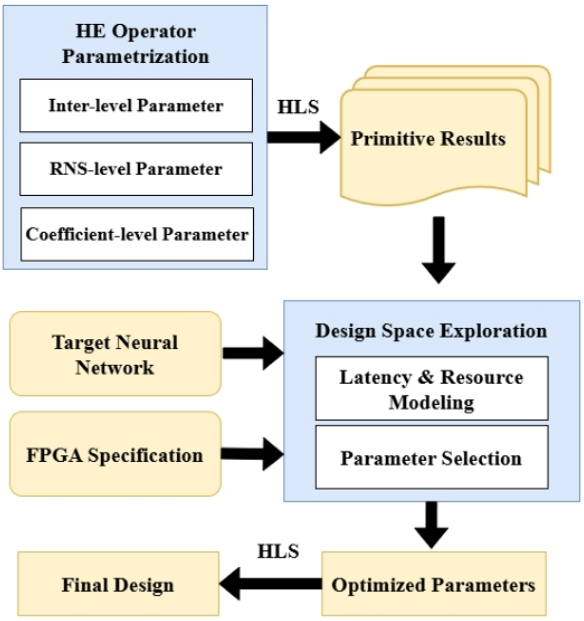
특히 2023 년 HPCA 에서 발표된 FxHENN 은 LoLa 기반 HE-CNN 모델을 FPGA 에서 가속하기 위해 암호화 파라미터부터 버퍼 구성, 자원 최적화까지 자동화 설계를 통해 latency, 에너지 효율, HE 기능 완전성 면에서 기존 선행 연구를 능가하는 성과를 달성했다고 소개

한다. PCmult, CCmult, Rescale, KeySwitch 같은 기본 HE 연산 모듈을 NTT, INTT, ModMul 등 하드웨어 기반 연산들에 의해 자동화된 DSE 설계로 구현함으로써 더 빠르고 더 효율적인 HE-CNN 추론을 가능하게 했다.

4. RNS level 변화를 고려한 설계 확장 필요성

한편 RNS-CKKS 구조에서 암호문 곱셈 연산마다 rescale 을 통해 RNS level 이 하나씩 감소하는데, 이는 곧 연산 가능 횟수의 감소를 말한다. 만약 추론 연산을 반복 실행하다 보면 네트워크 깊이가 level 의 크기보다 커서 level 이 0 에 도달하게 되는데, 여기서는 더 이상의 연산이 불가능하다. 이 경우 앞서 소개한 재부팅(bootstrapping)을 통해 level 을 초기화하는 과정이 필요한데, 이는 매우 느리고 비용이 가장 큰 연산이므로 latency 측면에서 비효율적이다. 그래서 대부분의 HE-CNN 논문[4][7]에서는 bootstrapping 을 피하기 위해 level 을 충분히 할당하고, 연산을 최적화하여 추론을 마무리할 수 있도록 설계한다. 대표적으로 FxHENN 을 보면, bootstrapping 없는 leveled HE 기반 CNN 가속을 제공한다고 설명할 수 있으나(bootstrapping 을 사용하지 않고도 일정 수준까지의 연산(depth)만 가능하도록 HE 파라미터(N, Q, L 등)를 미리 정해 놓은 구조), 그만큼 처리 가능한 depth 가 제한적이다.

따라서 RNS level 변화를 고려한 가변 구조를 통해 FHE-level 복잡성에도 유연하게 대응할 수 있도록 설계하는 아이디어를 제안하고자 한다. 다시 말해, 기존 FxHENN 처럼 고정된 level 을 가정하지 않고 다양한 level 상황을 고려해 모듈을 재활용하고 병렬도를 조절함으로써, FHE 환경에 더 근접한 확장성을 보여줄 수 있다는 것이다. 기존 HE-CNN 에서처럼 여러 층(layer)을 통과해야 하는데도 HE 연산 모듈이 고정된 level 수만 처리하도록 설계되었다면 어떤 layer 에서는 남은 level 이 부족해서 계산을 못하거나, 반대로 level 이 남아도는 모듈을 낭비할 수 있다. 그러므로 CNN 의 각 layer 마다 남은 level 이 다르다는 걸 디자인 타이밍에서 인지하고, HE 연산 모듈(PCmult, KeySwitch 등)의 병렬 구조와 메모리 사용량을 현재 level 에 맞게 조절(예를 들어, level 이 4 개인 ciphertext 는 NTT 4 개 코어 병렬로 처리, level 2 개인 건 NTT 2 개만 활성화하여 리소스 절약 + 성능 향상 달성)함으로써 bootstrapping 없이도 모든 layer 를 레벨 소진 전에 완료할 수 있도록 inference path 를 설계한다. 이를 통해 레벨 변화에 따라 연산 구조를 조정하는 동적 HE 연산 설계 기법을 제안하여, DSP 및 BRAM 의 활용도를 높이고 bootstrapping 없이도 전체 추론을 완료하도록 구현할 수 있다.



[그림 1] 새로운 디자인 프로세스

5. 결론

유럽의 GDPR 과 우리나라의 데이터 3 법을 지키면서 AI 기반의 4 차 산업 육성에 걸림돌이 되지 않으려면 동형암호와 같은 프라이버시를 온전히 보호할 수 있는 기술이 필요하다. 본 논문에서는 차세대 암호 기술인 동형암호와 이를 프라이버시 보존 기계학습에 적용한 HE-CNN 모델의 구현 및 최적화 기술 발전 흐름을 살펴보고, 나아가 동형암호의 하드웨어 가속기 특징을 고려할 때 level 을 고려한 설계 확장이 가능하다는 점에 착안해 새로운 디자인 프로세스를 제안하고자 했다. 이를 통해 HE-CNN 추론의 전 단계에 걸쳐 설계 유연성을 바탕으로 더 빠르고 더 효율적인 자원 분배를 달성할 수 있었다.

동형암호는 장단점이 명확한 차세대 보안(암호)기술로, 최적화 연구를 통한 성능 향상이 이루어진다면 정보보호 분야에서 유의미한 역할을 할 수 있을 것이다. 보안 문제의 딜레마인 보안 수준 및 비용과 성능 사이의 이상적인 trade-off 를 달성한 최신 암호 기술의 표준화를 통해 AI 기반의 실시간 통신에서 안전성을 확보할 수 있다면, 강력한 보안을 보장하는 것이 무엇보다 중요한 군사적 활용뿐만 아니라 상업적 활용에도 적극적인 어필을 할 수 있을 것으로 기대된다.

ACKNOWLEDGEMENT

이 논문은 2025 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행되었으며(No.RS-2023-00277326, 비수학적 접근법 중심의 암호화된 신경망 데이터 영오차 처리 기술), 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행되고(No.RS-2024-00438729, 익명화된 기밀실행을 이용한 전주

기적 데이터 프라이버시 보호 기술 개발), 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행되었으며(No.RS-2023-00277060, 개방형 엣지 AI 반도체 설계 및 SW 플랫폼 기술개발), 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구이며(No.2021-0-00528, 하드웨어 중심 신뢰계산기반과 분산 데이터보호박스를 위한 표준 프로토콜 개발), 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(IITP-2023-RS-2023-00256081). 그리고 2025 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었으며, 본 연구는 반도체 공동연구소 지원의 결과물임을 밝힙니다.

참고문헌

[1] Craig Gentry. “Fully Homomorphic Encryption Using Ideal Lattices.” Proceedings of the forty-first annual ACM symposium on Theory of computing, 2009.

[2] Cheon Jung Hee, Kim Andrey, Kim Miran, Song Yongsoo. “Homomorphic Encryption for Arithmetic of Approximate Numbers.” ASIACRYPT, 2017.

[3] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, J. Wernsing. “Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy.” International conference on machine learning. PMLR, 2016.

[4] A. Brutzkus, R. Gilad-Bachrach, O. Elisha. “Low latency privacy preserving inference.” International Conference on Machine Learning. PMLR, 2019.

[5] M. S. Riazi, K. Laine, B. Pelton, W. Dai. “Heax: An architecture for computing on encrypted data.” 25th International conference on architectural support for programming languages and operating systems, 2020.

[6] M. Han, Y. Zhu, Q. Lou, Z. Zhou, S. Guo, L. Ju. “coxhe: A software hardware co-design framework for fpga acceleration of homomorphic computation.” Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2022.

[7] Y. Zhu, X. Wang, L. Ju, S. Guo. “Fxhenn: Fpga-based acceleration framework for homomorphic encrypted cnn inference.” IEEE International Symposium on High-Performance Computer Architecture (HPCA). IEEE, 2023.

[8] Q. Lou, W.-j. Lu, C. Hong, L. Jiang. “Falcon: Fast spectral inference on encrypted data.” Advances in Neural Information Processing Systems, vol. 33, 2020.

[9] A. Al Badawi, C. Jin, J. Lin, C. F. Mun, S. J. Jie, B. H. M. Tan, X. Nan, K. M. M. Aung, V. R. Chandrasekhar. “Towards the alexnet moment for homomorphic encryption: Hcnn, the first homomorphic cnn on encrypted data with gpus.” IEEE Transactions on Emerging Topics in Computing, vol. 9, no. 3, 2020.