안전한 모델분리와 효율적인 SMPC 기반 PPML 기법

오재기, 최진아, 이수민, 윤병우, 신종호
LG 전자

jaeky.oh@lge.com, jina11.choi@lge.com, sum.lee@lge.com, byoungwoo.yoon@lge.com, jongho0.shin@lge.com

Secure model separation and efficient SMPC-based PPML method

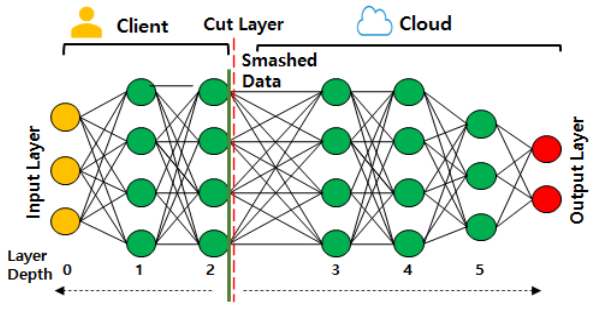
Jaeky Oh, Jina Choi, Sumin Lee, Byoungwoo Yoon, Jongho Shin
LG Electronics

요 약

본 논문은 Inversion attack 으로 부터 안전한 모델분리 방법과 모델정보를 보호하기 위한 SMPC 기반 PPML 기법을 기술하고 있다. Split-ML 은 원본 Data 가 그대로 노출되는 것을 막을 수 있지만, 모델분리의 위치에 따라 Inversion attack 에 취약하거나 혹은 Client 에게 과다한 모델 Layer 정보가 공개될 수 있다. 따라서 최적의 모델 분리 위치를 결정하는 방법과 모델 정보를 최대한 보호하면서도 공격(또는 inversion attack)에 강건한 효율적인 SMPC 사용법을 제시한다.

Ⅰ. 서 론

Split-ML 기술은 개인의 원본 data 를 노출하지 않고도 AI/ML 추론 서비스를 제공하거나 학습에 활용할 수 있도록 해 주는 기술이다[1]. <그림 1>은 원본 Data 대신 Client 에서 처리된 Smashed data 를 보내기 때문에 Data 를 보호하면서도 Cloud 의 부담을 줄일 수 있다.

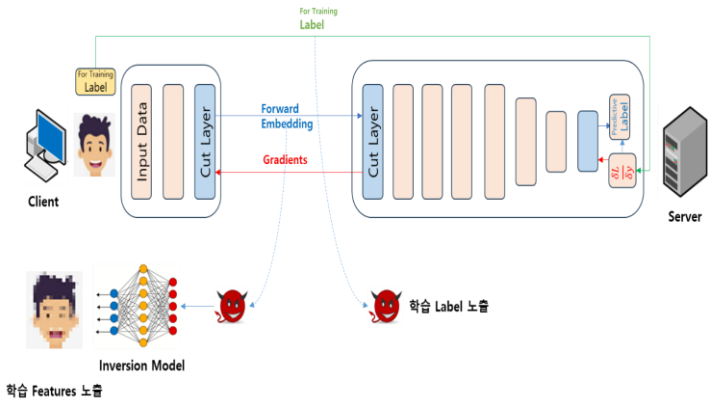


<그림 1> Split-ML 의 구조

그러나 Client 의 처리 영역을 지나치게 작게 설정하면 Smashed data 의 추상화 정도가 낮아져 Inversion attack 에 취약할 위험이 있다[2].

<그림 2>는 Split-ML 에서 각 Cut layer 위치에 따른 Smashed data 로 복원된 이미지를 보여준다. 가장 왼쪽의 이미지는 실제 Input data 로 들어오는 이미지이며, Layer depth 가 증가할수록 복원 이미지는 흐릿하고, 대략적으로 변하기 때문에 원본 이미지를 구분하기가 어려워진다.

<그림 2> Inversion Attack 으로 복원된 이미지



<그림 3> Split ML(Vanila Structure) 모델의 취약점

<그림 3>은 Vanilla Split-ML 구조를 보여 주는데, 이 경우 학습 Features 와 Label 이 노출될 위험이 있다. 학습 Label 의 노출은 Loss 계산이 Client 에서 이루어지는 U-shape 구조로 완화할 수 있지만, 잘 훈련된 Inversion model 을 사용한 공격으로부터 여전히 노출 위험성을 가진다.

Ⅱ. 본론

Inversion attack 으로부터 안전한 최소 Layer depth 를 Secure layer 로 정의할 수 있다. 이는 Split-ML 의 Cut layer 로 직접 사용되거나, 혹은 효율적인 SMPC 구현에 사용된다.[5] Split-ML 에서 안전한 Layer depth 를 결정하는 것은 중요하며, 안정성은 Target 과 Inversion 한 결과와의 차이를 나타내는 Inversion metrics 로 확인할 수 있다. Input data 가 이미지일 경우에는 PSNR, LPIPS, MSE 가 사용되며, MSE 식은 다음과 같다[3].

$$MSE = \frac{1}{HW}\sum_{i=0}^{H}\sum_{j=0}^{W}[T(i,j) - I(i,j)]^2$$

T(i,j)는 (i,j) 좌표에서의 Target Image 값이며 I(i,j)는 Inversion 모델을 사용하여 복구한 Image 값이다. H 와 W 는 Image 의 높이와 넓이다. PSNR(Peak Signal to Noise Ratio)는 MSE 를 사용하여 아래와 같이 표현된다.

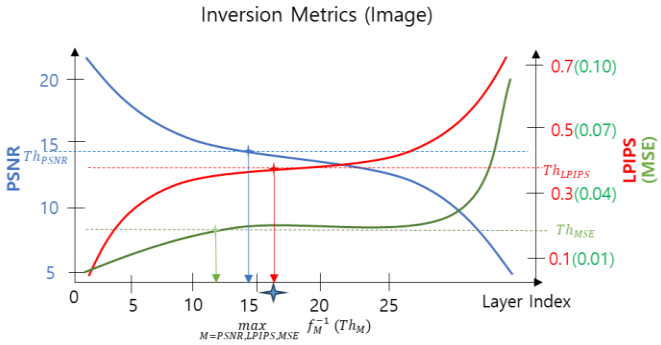
$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_T^2}{MSE}\right), \quad MAX_T = 2^B - 1$$

,

LPIPS(Learned Perceptual Image Patch Similarity)는 이미지의 유사도를 사람의 인식 기반으로 표현한 지표이다. 이는 2 개 이미지에서 Pretrain(ImageNet)된 모델 Layer l 에서의 activation map 을 비교하여 아래식과 같이 표현된다.

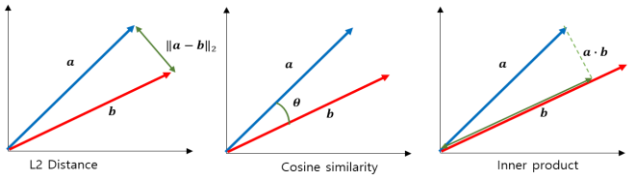
$$LIPIS = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w^l \odot \left(\hat{y}^l_{hw} - \hat{y}^l_{0hw} \right) \right\|_2^2$$

<그림 4>는 Layer 별 Inversion metrics 에 따른 Secure layer 결정방법 중 하나를 보여주고 있다[4]. 그림에서 각 그래프는 Metrics 별로 Layer depth 에 따른 Inversion metrics 값(f_M)을 표시한다. 상한 또는 하한 Inversion metrics 기준값(Th_{PSNR}, Th_{LPIPS}, Th_{MSE})이 선정된 경우라면, 그래프를 활용하여 Secure layer 를 결정할 수 있다. 이때 각 Metrics 의 기준값들은 사용 모델이나 이미지에 따라 달라질 수 있으며, 정량 혹은 정성평가로 결정될 수 있다.



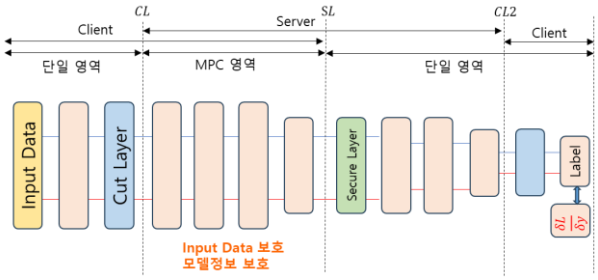
<그림 4. Secure Layer 결정법>

Input data 가 Image 가 아닌 다른 Data 의 경우에는 Vector Similarity 이용하여 Inversion metrics 를 측정할 수 있으며 <그림 5>는 Vector similarity 를 위한 Metrics 을 보여주고 있다.



<그림 5. Vector similarity 를 위한 Metrics>

Split-ML 이 Inversion attack 에 안전 하려면 앞에서 결정된 Secure layer 까지 Client 에서 처리하고 그 때의 Smashed data 를 Cloud 에게 전송해서 나머지 ML 을 수행하면 된다. 하지만 Cloud 는 Secure layer 까지 모델정보를 Client 에게 노출하게 되며, 특히 Server 의 중요한 모델정보가 있는 경우에는 문제가 심각해진다. 이를 해결하기 위해서, 우리는 PETs 기술인 SMPC 를 사용하여 Server 의 모델 정보를 노출하지 않고도 ML 을 동작 시킬 수 있다. 하지만 모델전체에 SMPC 기술을 적용하기에는 많은 통신부담을 발생시킨다. 따라서 제안된 방식에서는 Inversion attack 에 안전한 Secure layer 까지만 SMPC 로 모델 정보를 보호하고, 이후부터는 일반 ML 을 사용하여 통신 부담을 최소화하였다.



<그림 6. 모델정보 보호기능이 추가된 효율적 SMPC>

<그림 6>은 제안된 방식의 구조를 보여준다. Client 에서 Input data 로부터 Cut layer(*CL*)까지 평문 영역에서 실행하고 *CL* 부터 Secure layer(*SL*)까지 SMPC 를 통해서 연산을 수행한다. SMPC 는 Client 와 Server 가 각각의 비밀조각으로 같은 연산들을 수행하는 것이다. SMPC 영역인 *CL*~*SL* 사이에서 Client 는 Server 의 해당영역 모델정보를 알지 못하고, Server 는 Client 에서 생성된 Embedding 을 알지 못하게 된다. *SL* 에서 Server 는 Client 로부터 MPC 연산결과를 전달받아 완전한 Embedding 을 복구한 후, *SL*~*CL2* 까지 평문 영역으로 연산한다. 마지막으로 Client 는 *CL2* 에서 Server 에서 연산된 Embedding 을 이용하여 최종 예측 Label 을 계산할 수 있다. 계산값과 학습 Label 를 이용하여 Loss 의 Gradient 인 $\frac{\delta L}{\delta y}$ 을 계산할 수 있고 앞서 설명한 Forward Propagation 의 반대방향으로 Backward Propagation 을 수행하여 학습을 수행한다.

Ⅲ. 결론

본 논문에서는 Split-ML 과 SMPC 를 동시에 적용하여 입력데이터와 모델정보를 모두 보호할 수 있는 방안을 제시하였다. 제안된 기술에서는 Layer depth 별로 안정성을 수치화한 Inversion metrics 값을 측정하고 요구되는 Threshold 값과의 비교를 통해 Secure layer 를 결정하였다. 또한 모델의 정보 노출을 막을 수 있는 Cut layer 를 설정하고 Cut layer 와 Secure layer 구간만을 SMPC 로 보호하여, 통신 부담을 최소화한 효율적인 방식을 제시하였다.

참 고 문 헌

[1] Yoshitomo M. “Splitting distilled deep neural networks”, IEEE Access, 8:212177- 212193, 2020

[2] Sanjeev A. ” Why are deep nets reversible”, ICLR, 2015.

[3] Richard Z. “The unreasonable effectiveness of deep features as a perceptual metric”, CVPR, 2018.

[4] Xin D. “Privacy Vulnerability of Split Computing to Data-Free Model Inversion Attacks”, BMVC, 2022

[5] A. B. Alexandru, “Cloud-based mpc with encrypted data,”, IEEE CDC, 2018