

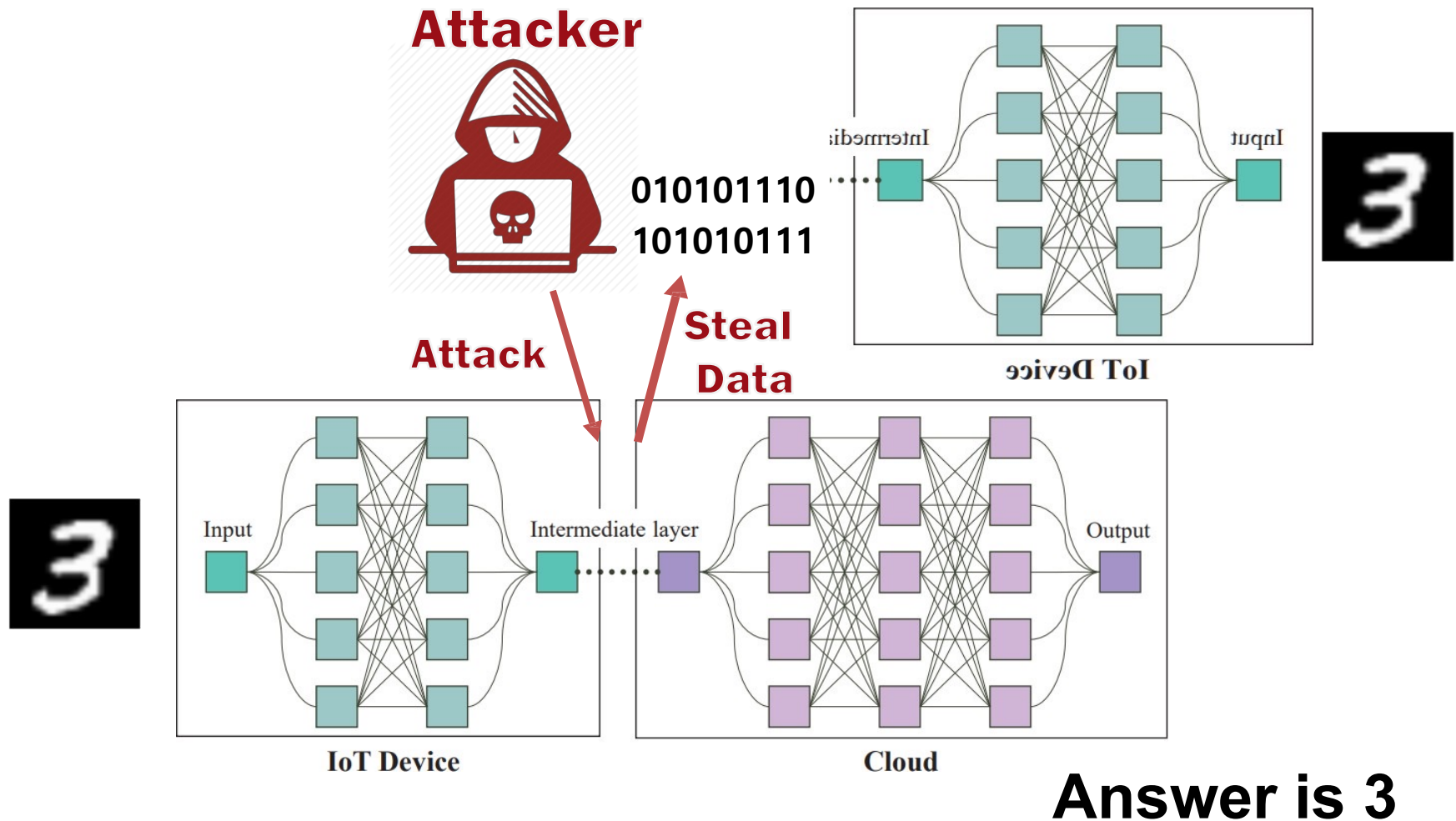
Can differential privacy practically protect collaborative deep learning inference for IoT?

Jihyeon Ryu

2023.11.3

Kwangwoon University
School of Computer and Information Engineering
Cryptography & Cyber Security Lab

Overview



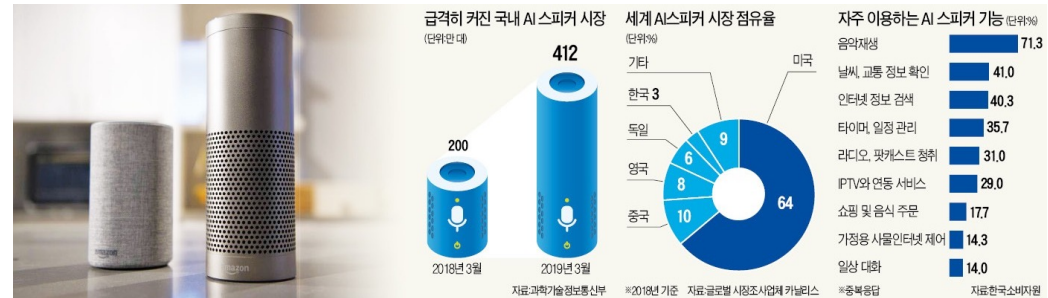
Introduction (Motivation)

[글로벌] 페이스북, 사생활 침해 논란 얼굴인식 프로그램 사용 중단키로

▲ 조만수 기자 | © 임력 2021.11.03 16:25 | 댓글 0



내가 한 말 녹음해 음성인식을 강화?...AI 스피커 '사생활 침해' 논란



- 인공지능은 학습 과정에서 대량의 데이터가 사용되고 있어, 수집한 데이터에 대한 **정보보호 관련 문제가 제기되고 있음**
 - Facebook은 2020년 7월, 개인정보 보호법 위반에 6억 5천만달러의 합의금을 지불, **사생활 침해 논란으로 얼굴인식 프로그램을 사용 중단**하기로 함
 - 인공지능 스피커의 음성 인식 수준을 높이기 위해 **이용자의 목소리를 수집**하며 사생활 침해 논란이 일어남
 - 2021년 1월, 인공지능 챗봇 '이루다'는 학습과정에서 수집한 60만명의 문장에서 데이터 보호 조치가 빈약하여 **특정인의 실명과 주소 등의 개인정보가 유출**됨

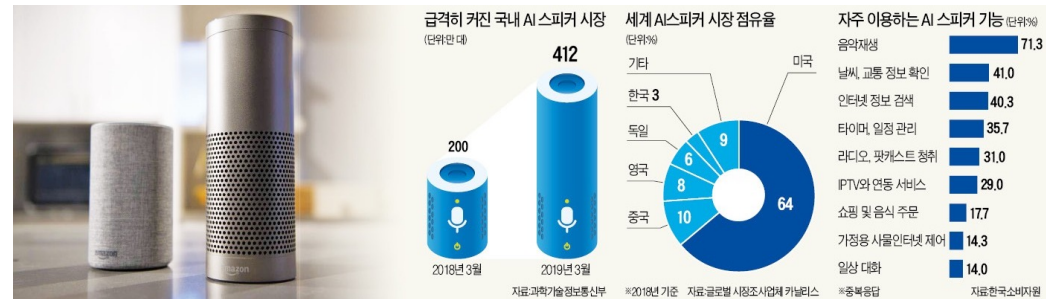
Introduction (Motivation)

[글로벌] 페이스북, 사생활 침해 논란 얼굴인식 프로그램 사용 중단키로

▲ 조만수 기자 | © 임력 2021.11.03 16:25 | 댓글 0

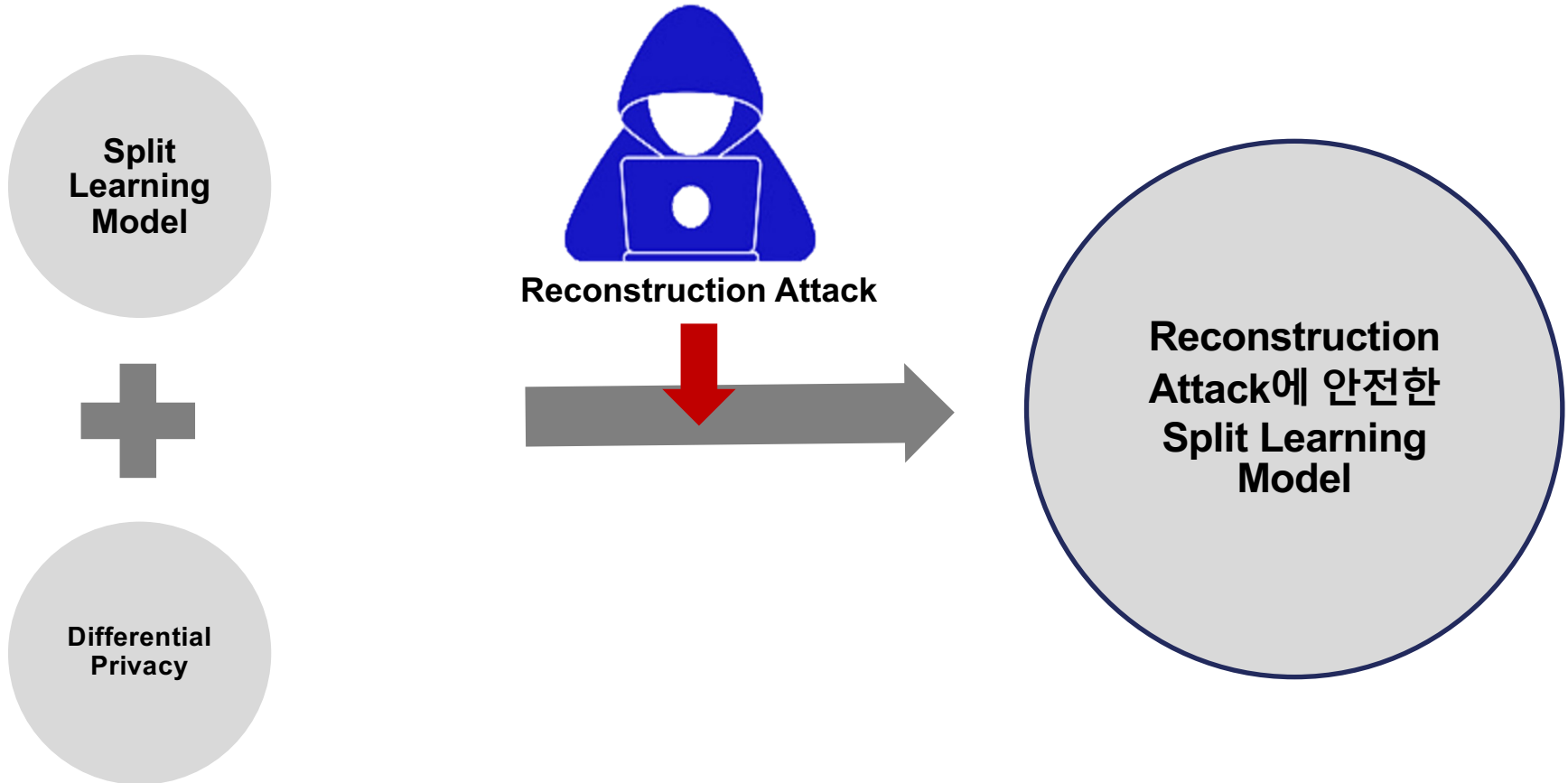


내가 한 말 녹음해 음성인식을 강화?...AI 스피커 '사생활 침해' 논란



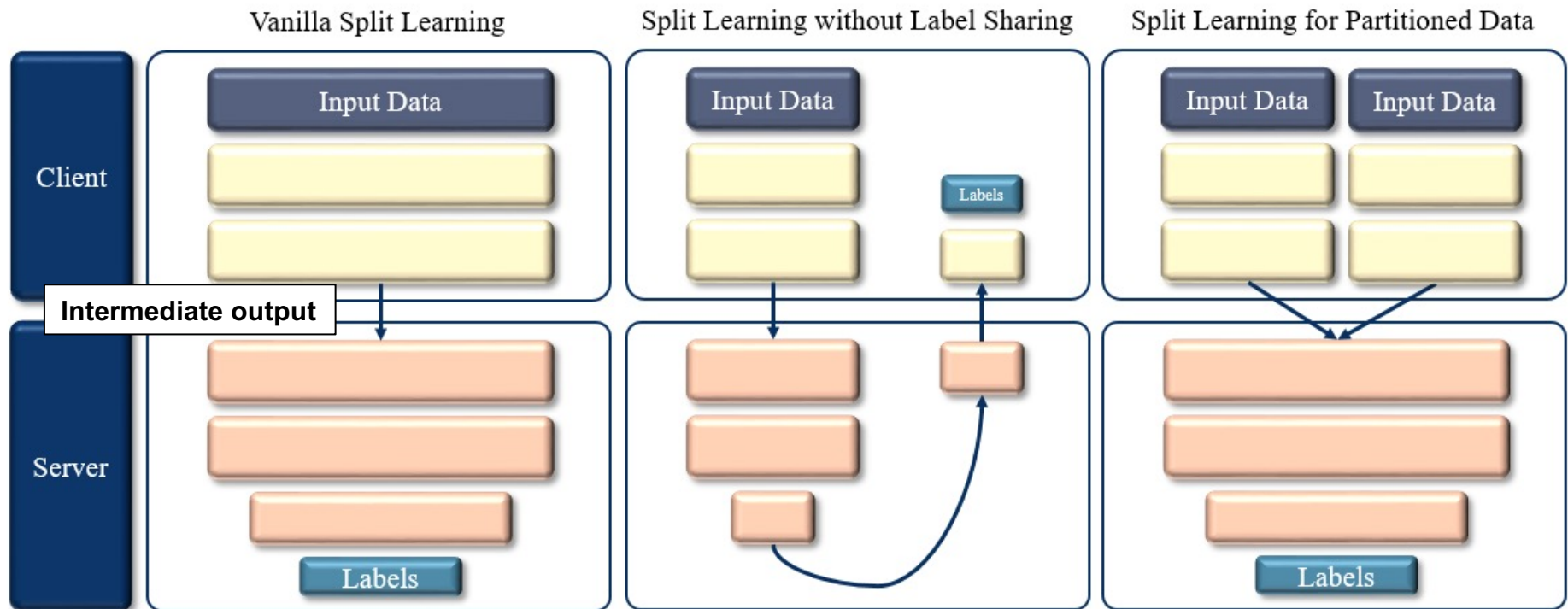
- 인공지능 학습을 위해 수집한 데이터 및 딥러닝 모델을 사용하기 위해 이용자가 제공한 데이터를 보호하기 위해, **데이터에 대한 보호 조치**가 필요함
- 학습 데이터를 위한 보호를 위해 **Split Learning**이 제안됨

Introduction (Motivation)

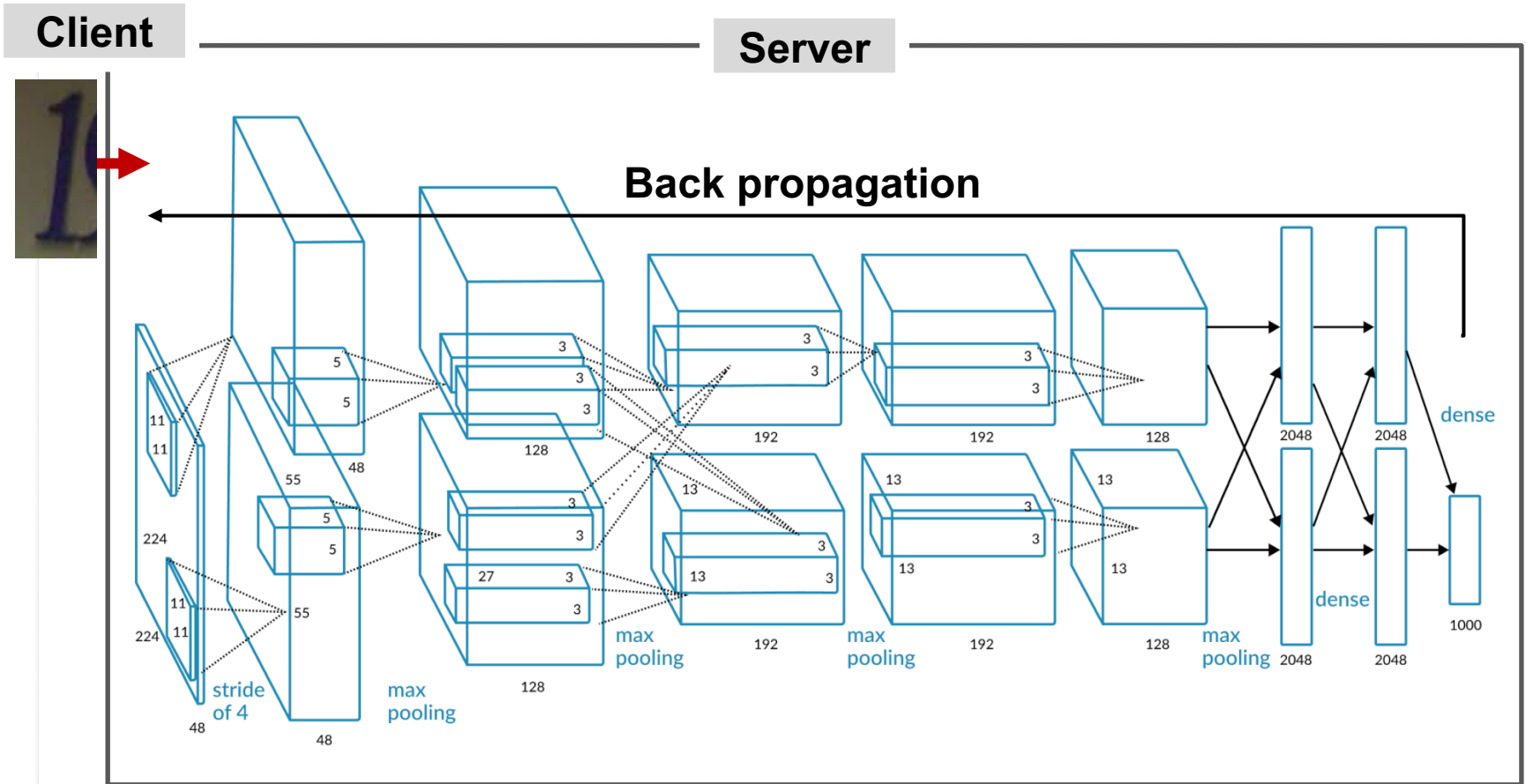


Background (Split Learning)

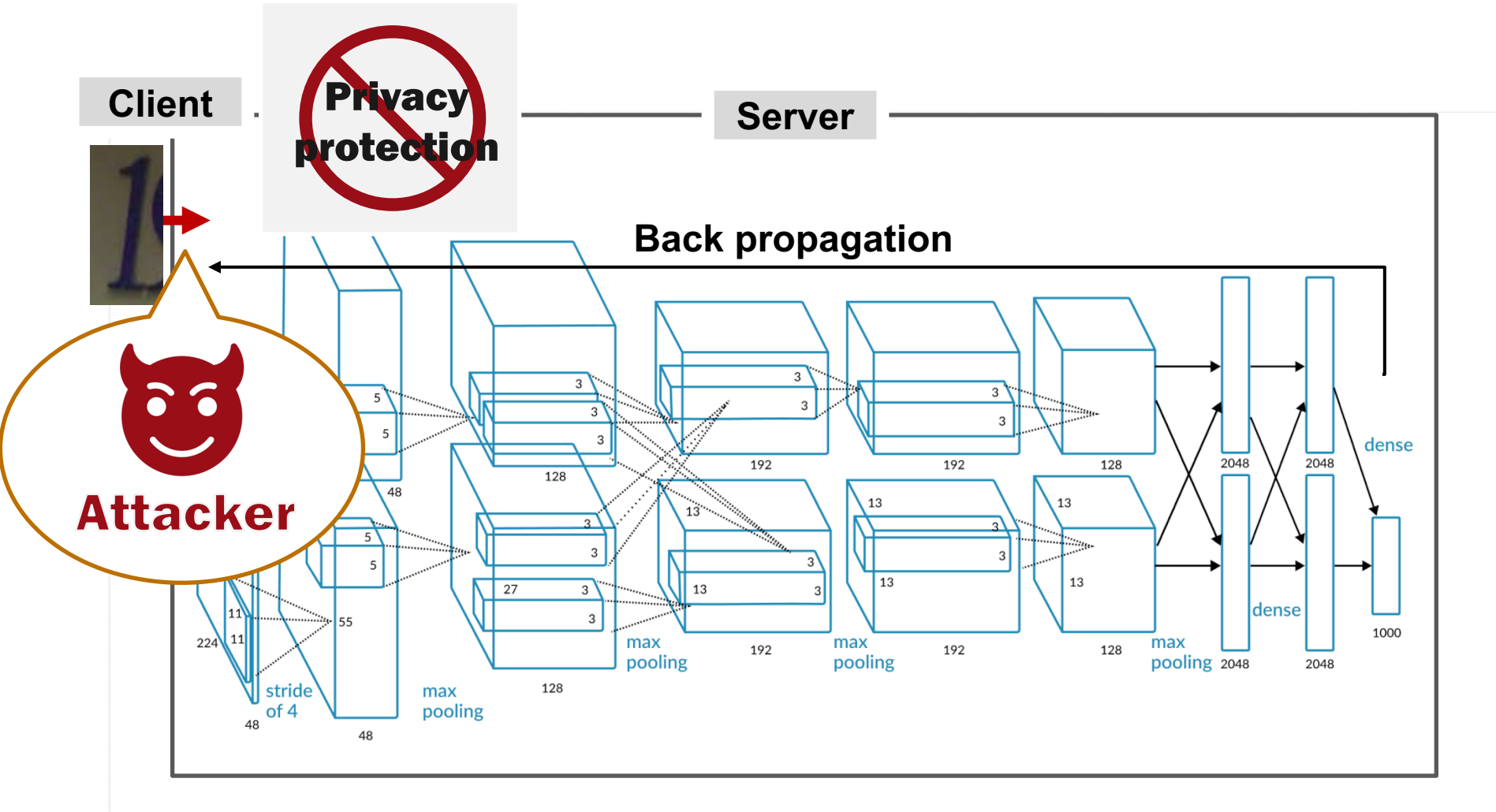
- Split Learning



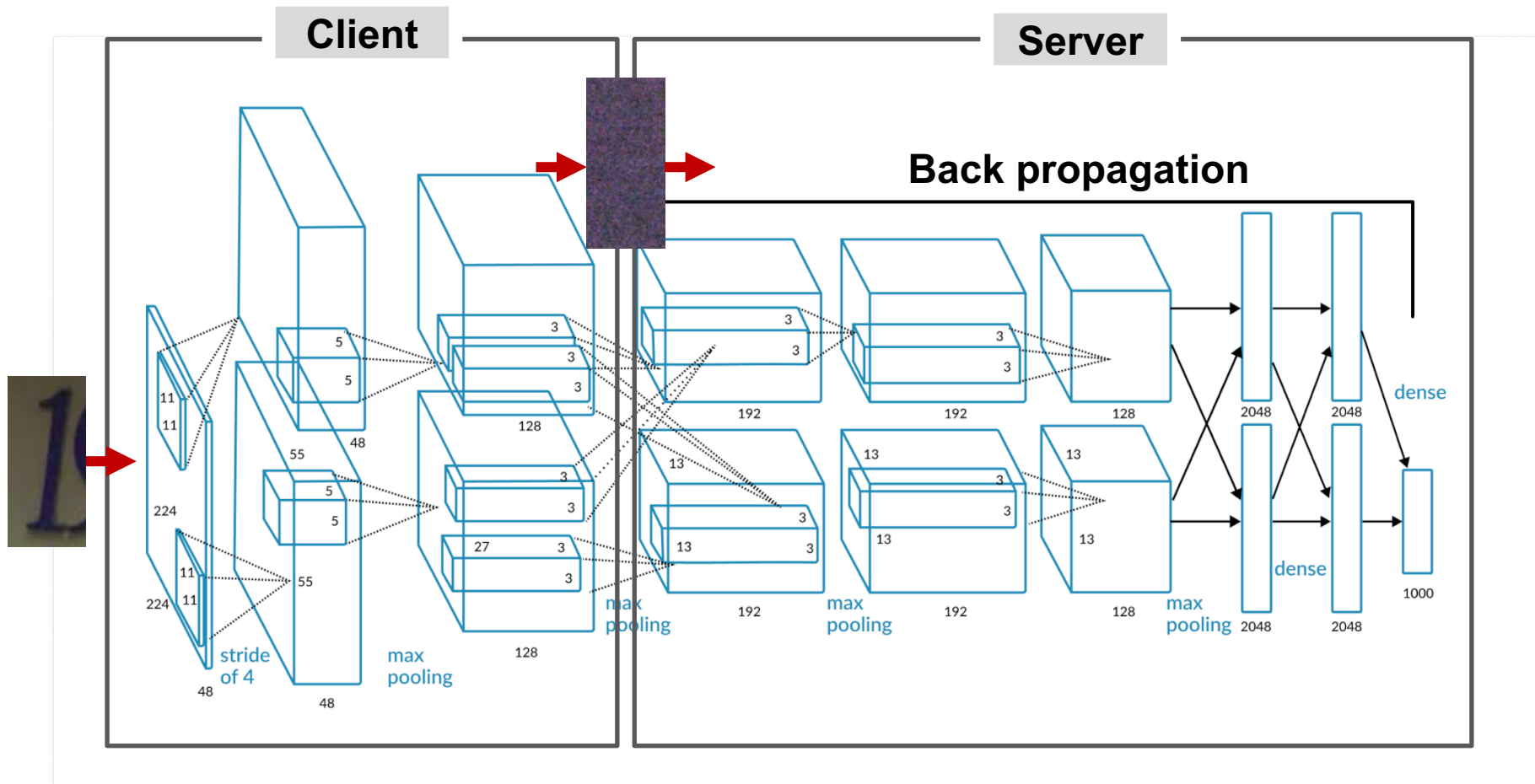
Neural Networks



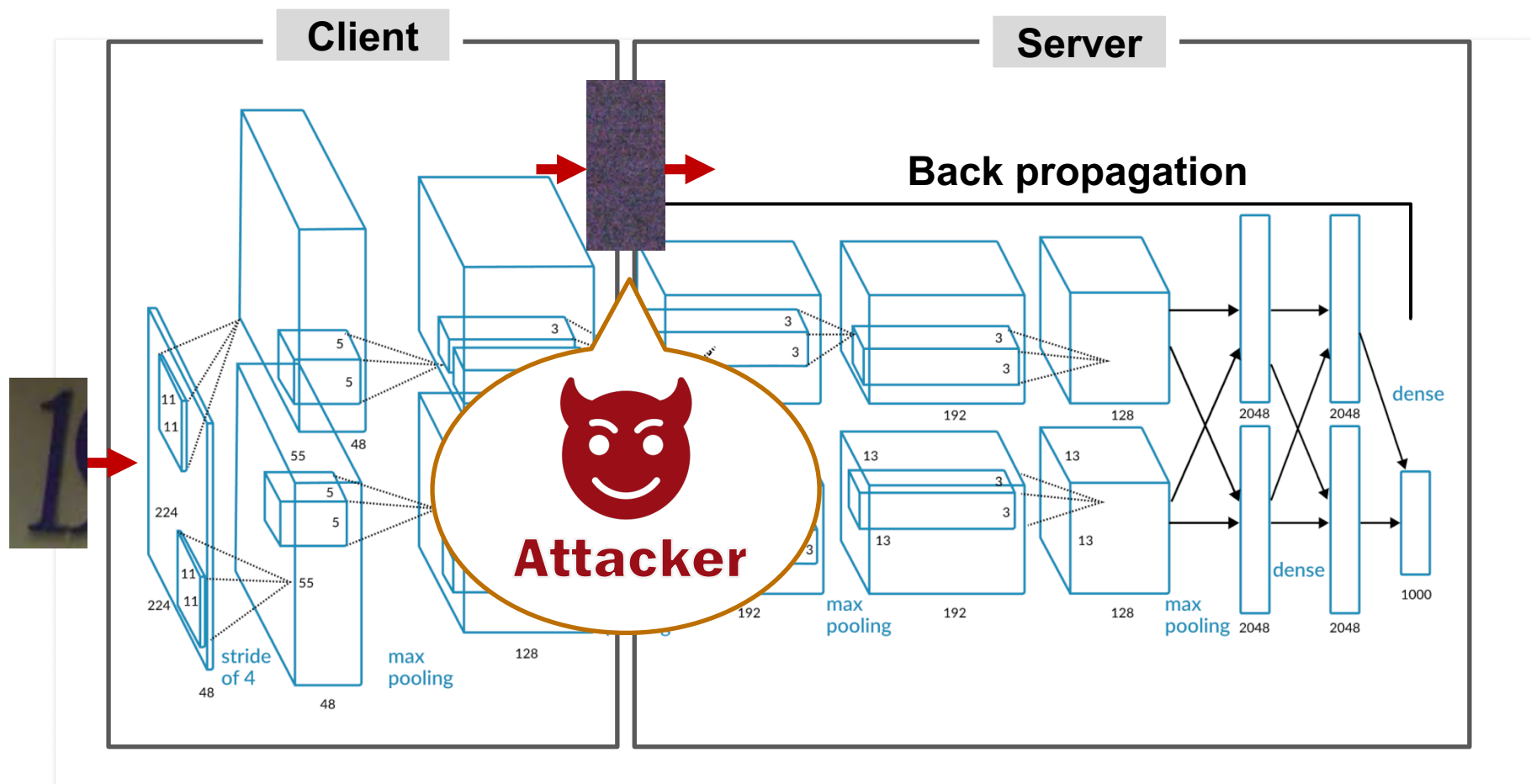
Attack on Neural Networks



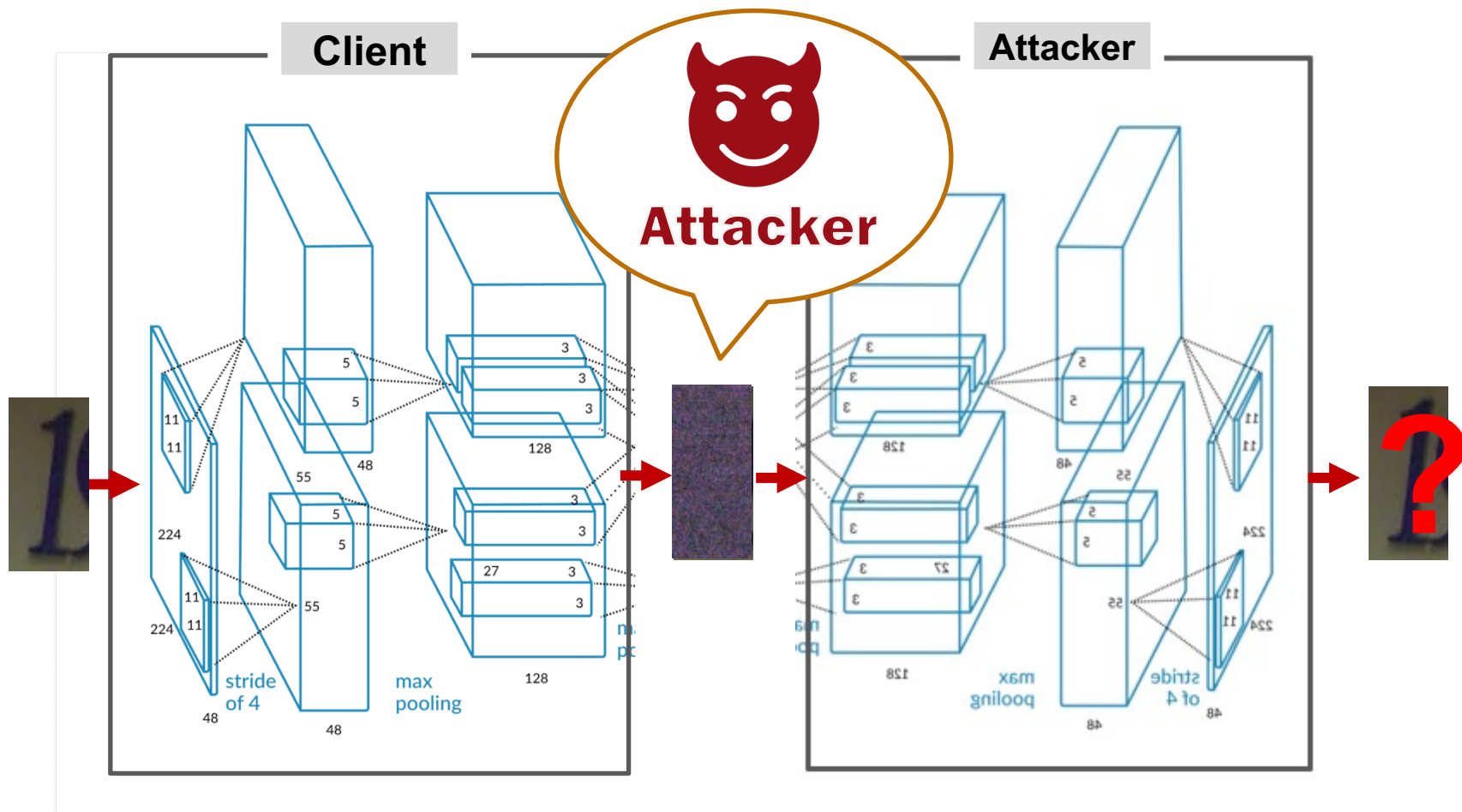
Split Learning



Attack on Split Learning



Reconstruction Attack



Background (Differential Privacy)

- Differential Privacy

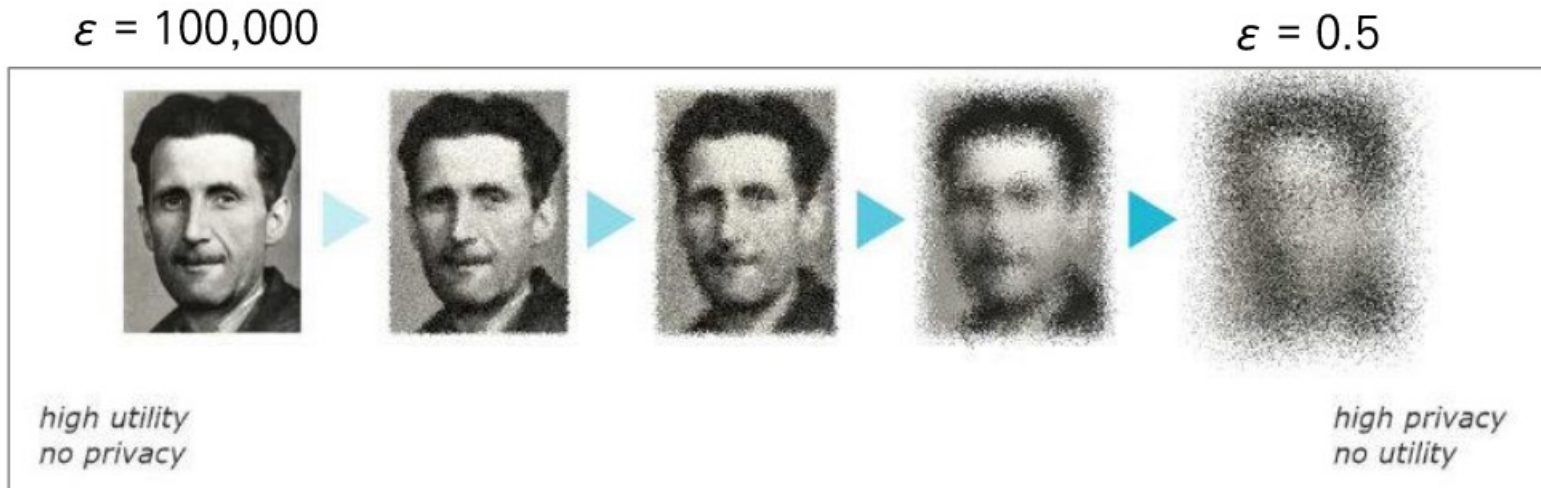
Define that there is a neighbor database when two databases have the same code except for one record t .

Definition 1. Given any two neighboring inputs D and D' which differ in only one data item, a mechanism M provides ϵ -differential privacy if $\Pr[M(D) \in S] \leq e^\epsilon \cdot \Pr[M(D') \in S]$.

- ϵ 이 0일 경우, 두 인접 데이터베이스에서 동일한 통계 결과가 나올 확률이 같음
 - 공격자는 통계 결과에서 절대 특정 레코드를 추론할 수 없음
- ϵ 이 무한대로 커질 경우, 두 인접 데이터베이스에서 동일한 통계 결과가 나올 확률의 차이가 커짐
 - 공격자는 통계 결과에서 특정 레코드를 추론하기 쉬움
- ϵ 값이 작아지면 프라이버시 정도가 높아지나 활용 가능성이 낮아지는 것이며, ϵ 값이 커지면 프라이버시 정도가 낮아지나 활용 가능성이 높아짐

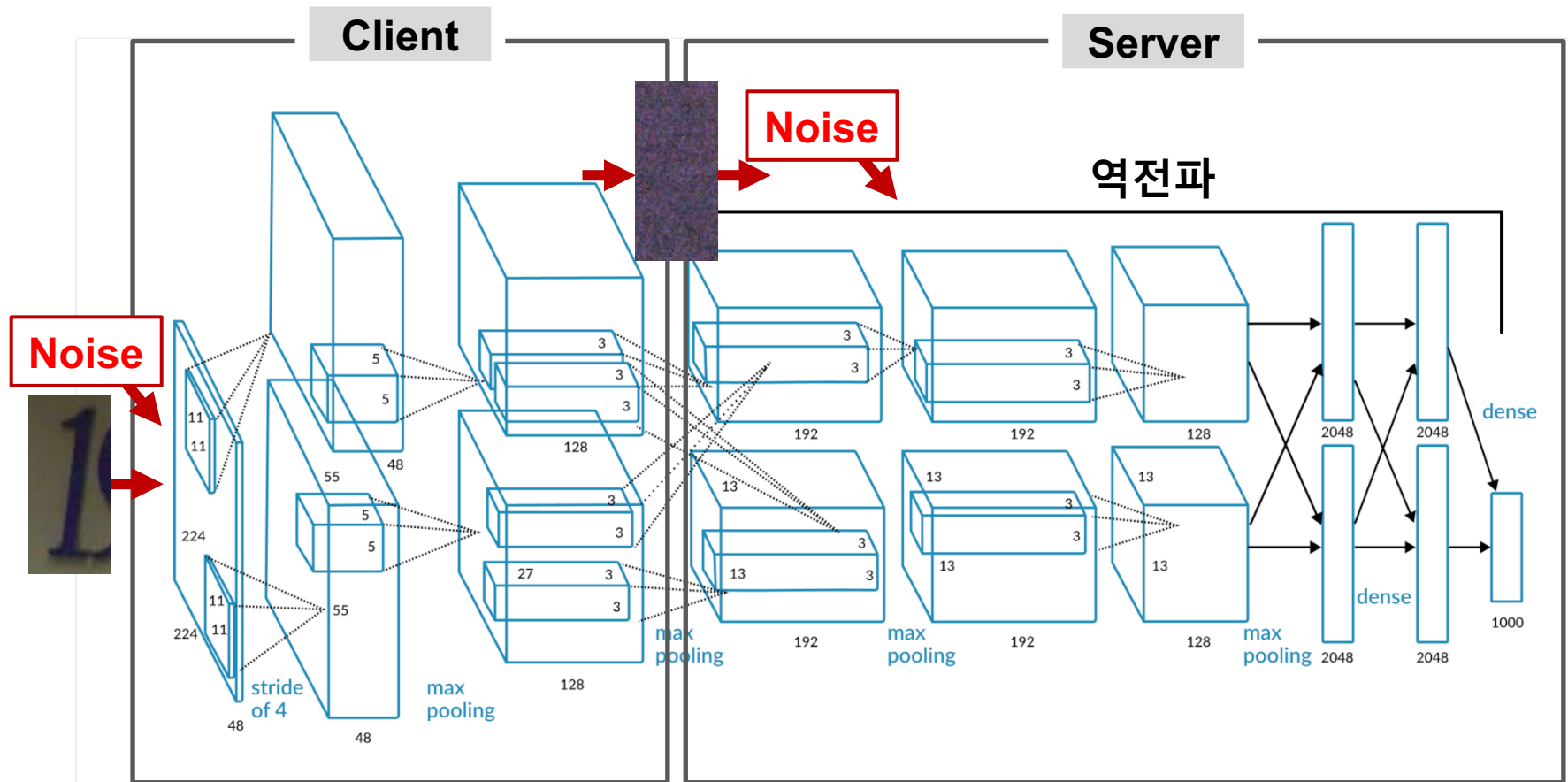
Background (Differential Privacy)

- Differential Privacy

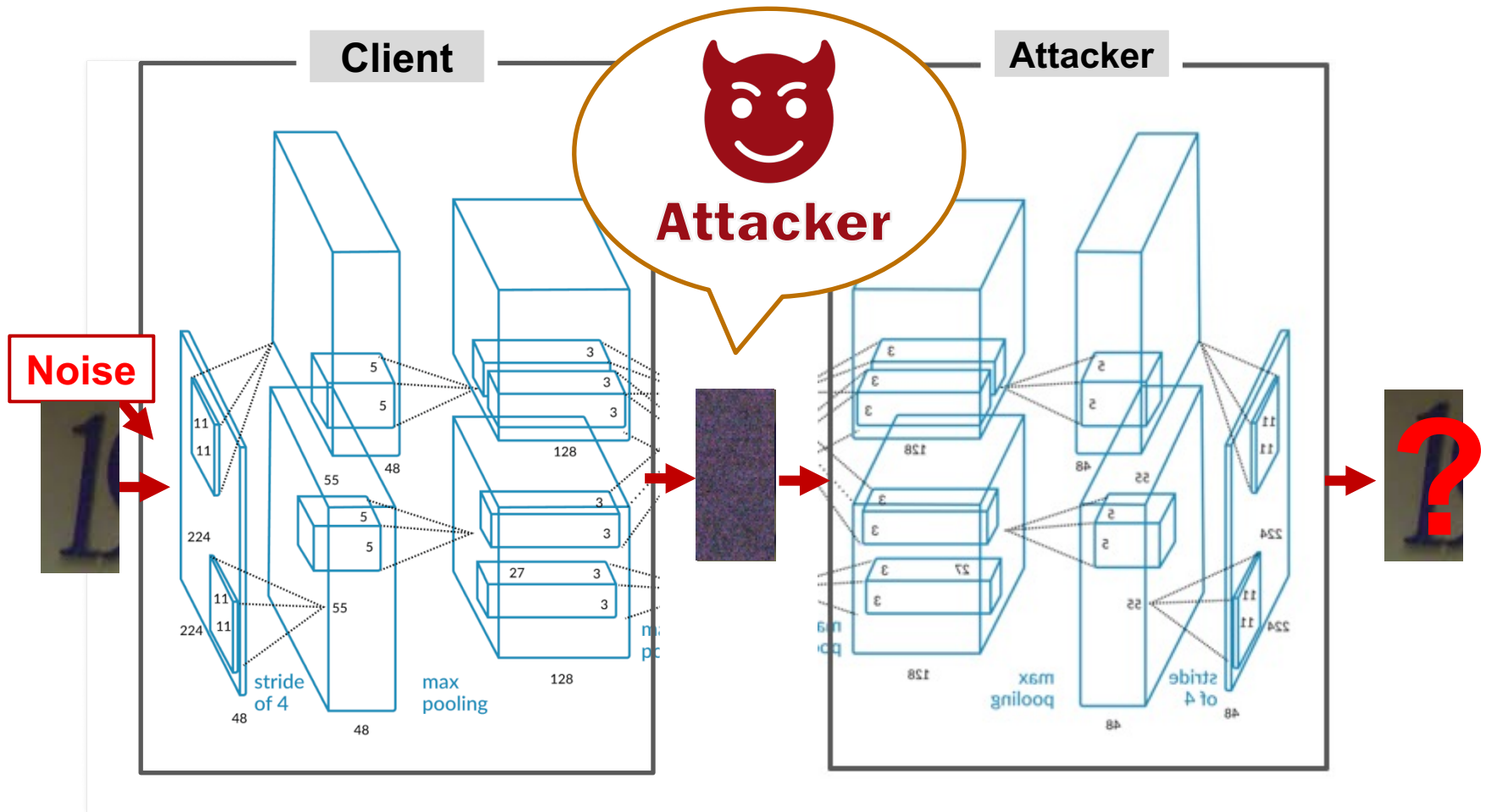


- ϵ 이 0일 경우, 두 인접 데이터베이스에서 동일한 통계 결과가 나올 확률이 같음
 - 공격자는 통계 결과에서 절대 특정 레코드를 추론할 수 없음
- ϵ 이 무한대로 커질 경우, 두 인접 데이터베이스에서 동일한 통계 결과가 나올 확률의 차이가 커짐
 - 공격자는 통계 결과에서 특정 레코드를 추론하기 쉬움
- ϵ 값이 작아지면 프라이버시 정도가 높아지나 활용 가능성이 낮아지는 것이며,
 ϵ 값이 커지면 프라이버시 정도가 낮아지나 활용 가능성이 높아짐

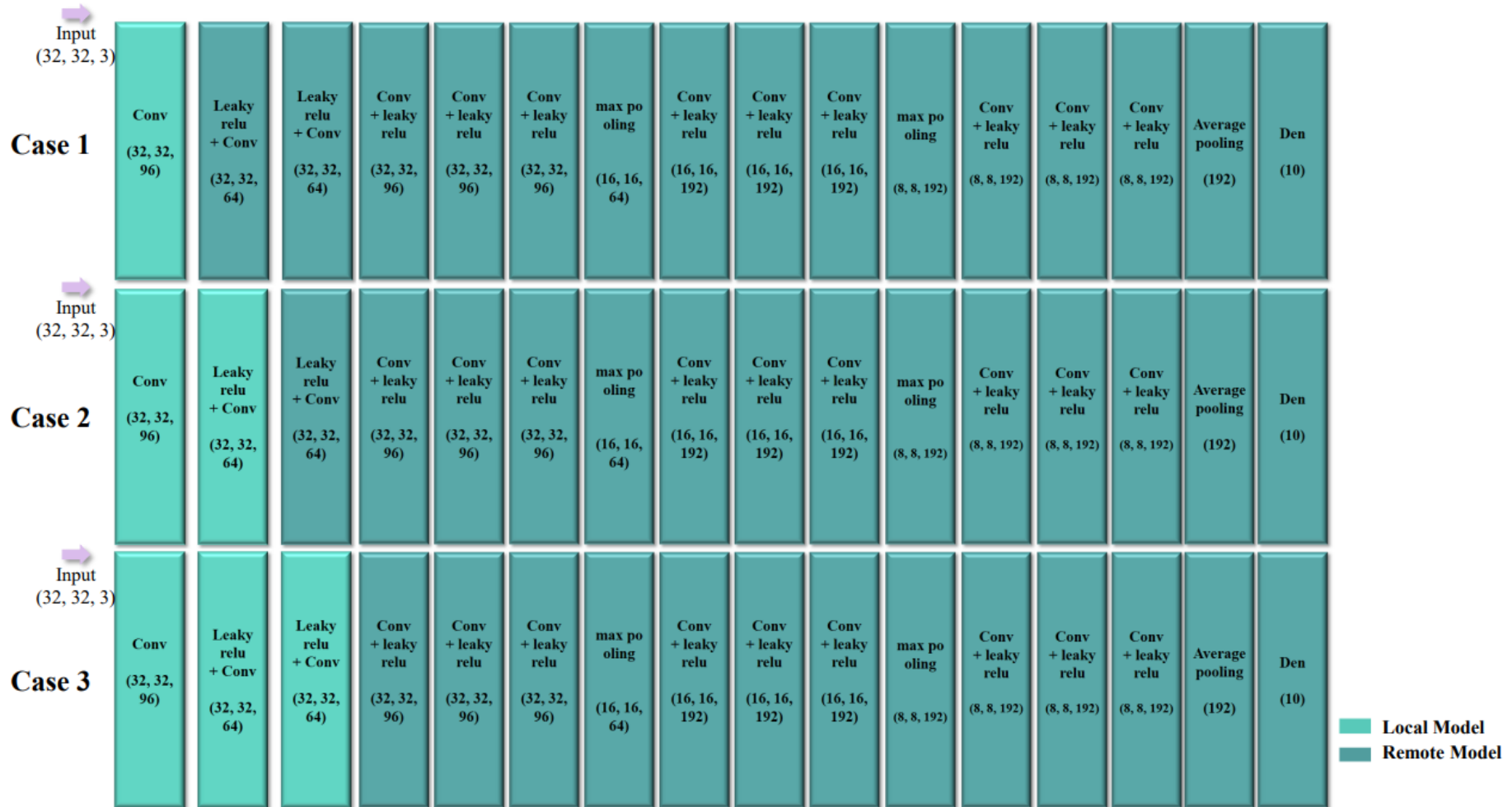
Split Learning with Differential Privacy



Reconstruction Attack on SL with DP



Models



Quantitative Metrics

MSE (Mean Square Error)

$$MSE(A, B) = \frac{1}{m \cdot n} \sum_{i,j=1}^{m,n} \|A(i, j) - B(i, j)\|^2$$

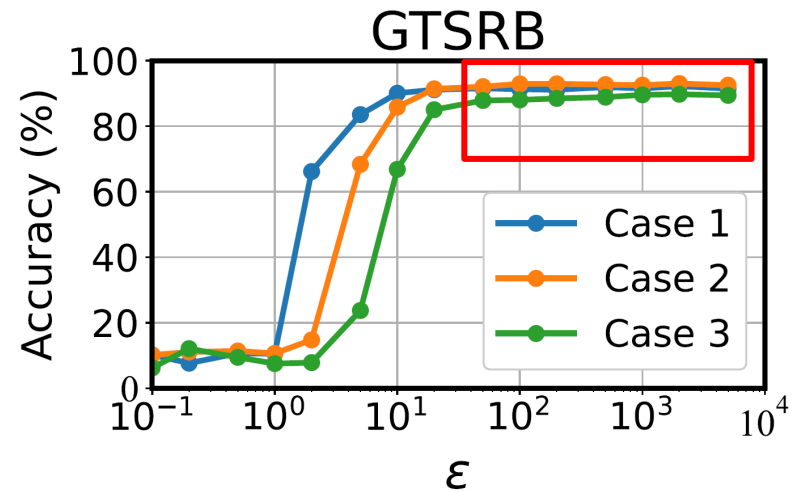
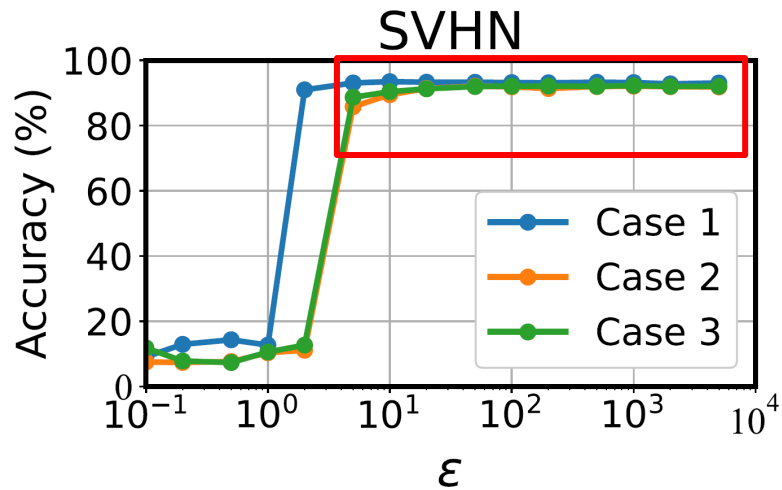
SSIM (Structural Similarity)

$$SSIM(A, B) = \frac{(2\mu_A\mu_B + C_1)(2\sigma_{AB} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_A^2 + \sigma_B^2 + C_2)}$$

Peak Signal to Noise Ratio (PSNR)

$$PSNR(A, B) = 10\log_{10}\left(\frac{255^2}{MSE(A, B)}\right)$$

Accuracy



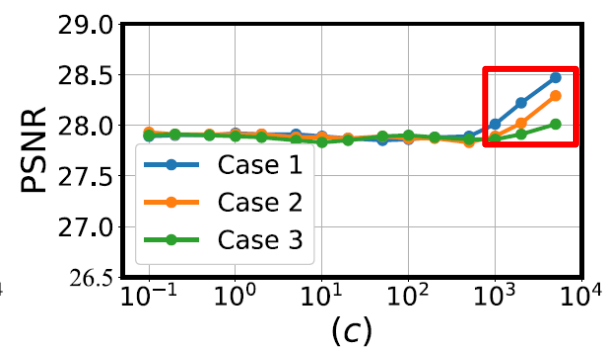
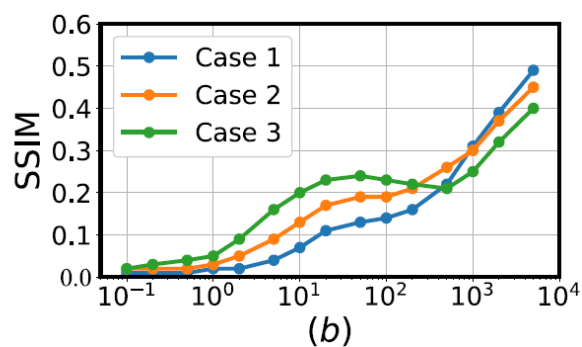
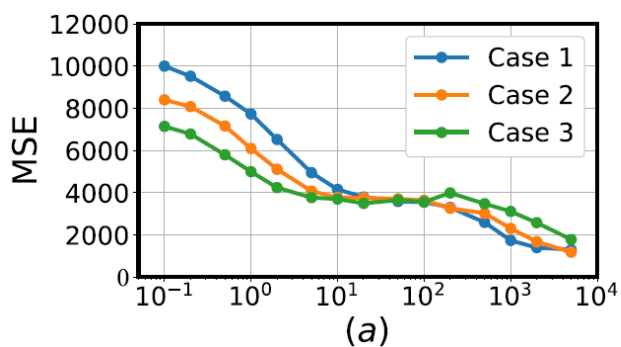
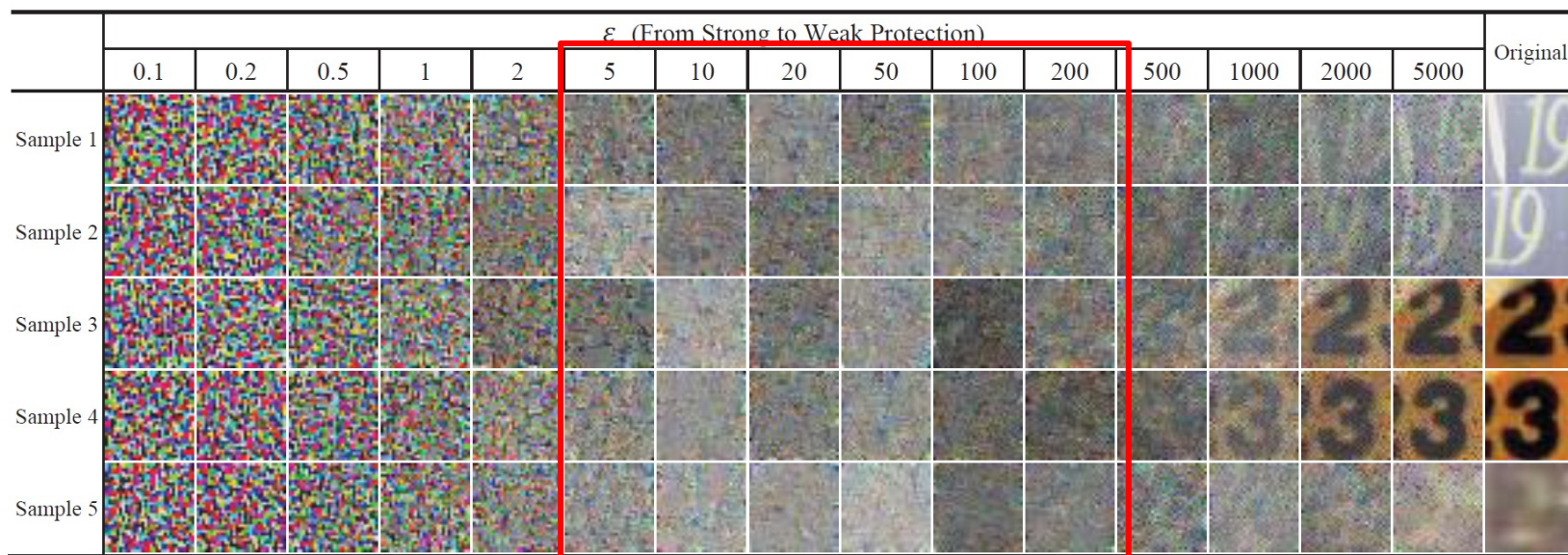
Model	SVHN	GTSRB	STL-10	CIFAR-10
Case 1	93.498%	92.676%	69.576%	82.932%
Case 2	92.695%	95.284%	67.448%	77.795%
Case 3	92.953%	92.869%	67.333%	84.500%
Average	93.049%	93.610%	68.119%	81.742%

SVHN Attacking Results

	ϵ (From Strong to Weak Protection)															Original
	0.1	0.2	0.5	1	2	5	10	20	50	100	200	500	1000	2000	5000	
Sample 1																
Sample 2																
Sample 3																
Sample 4																
Sample 5																

	ϵ (From Strong to Weak Protection)															Original
	0.1	0.2	0.5	1	2	5	10	20	50	100	200	500	1000	2000	5000	
Sample 1																
Sample 2																
Sample 3																
Sample 4																
Sample 5																

SVHN Attacking Results



GTSRB Attacking Results

	ϵ (From Strong to Weak Protection)															Original
	0.1	0.2	0.5	1	2	5	10	20	50	100	200	500	1000	2000	5000	
Sample 1																
Sample 2																
Sample 3																
Sample 4																
Sample 5																

	ϵ (From Strong to Weak Protection)															Original
	0.1	0.2	0.5	1	2	5	10	20	50	100	200	500	1000	2000	5000	
Sample 1																
Sample 2																
Sample 3																
Sample 4																
Sample 5																

GTSRB Attacking Results

