R-CONV: AN ANALYTICAL APPROACH FOR EFFICIENT DATA RECONSTRUCTION VIA CONVOLUTIONAL GRADIENTS

TAMER AHMED ELTARAS*, QUTAIBAH MALLUHI+, ALESSANDRO SAVINO*, STEFANO DI CARLO*, ADNAN QAYYUM+

* Politecnico di Torino, + Qatar University

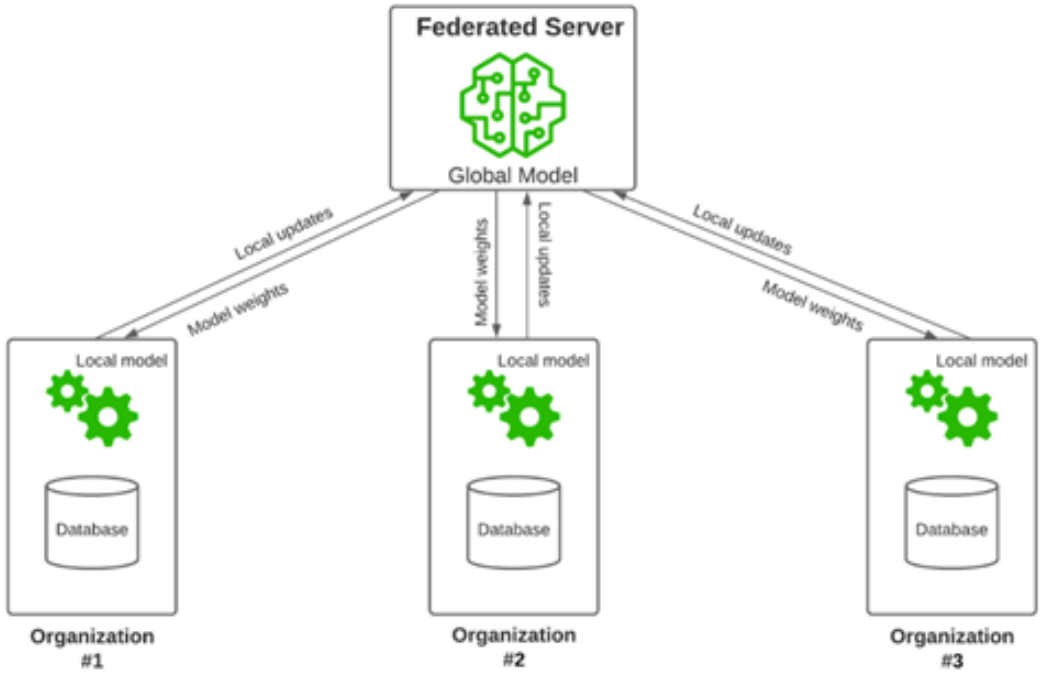
This study is partially supported by: the “COLTRANE-V” project - funded by the Ministero dell’Universit` a e della Ricerca within the PRIN 2022 program (D.D.104 - 02/02/2022). and SERICS project (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

DATA IN ML

- The performance and robustness of the machine learning models rely on the access to large datasets of good quality.
- A conventional approach is to gather all data at a central server and use it to train the model.
- Such datasets usually include privacy-sensitive information.
  - concerns about data privacy.
  - leaving a lot of valuable data inaccessible.

FEDERATED LEARNING

- Decentralized approach to training machine learning models.
- Doesn't require an exchange of data from client devices to global servers.
- The raw data remain locally.
- The final model is formed by aggregating the local updates.



PRIVACY & SECURITY ISSUES

- Many attacks have shown the vulnerability of federated learning systems:
  - Inversion attacks
  - Membership inference attacks
  - Poisoning attacks

PRIVACY PRESERVING TECHNIQUES

- Several privacy-preserving techniques have been introduced to mitigate these risks.
- The main challenge is to balance privacy, security, and efficiency in federated systems.
- There is a growing need for systems that can simultaneously address multiple attack types.
- To achieve the goal of identifying data leakage sources towards developing comprehensive solutions.
  - Developed an Analytical Approach for Efficient Data Reconstruction via Convolutional Gradients.

PROBLEM STATEMET

- Initial Setup:

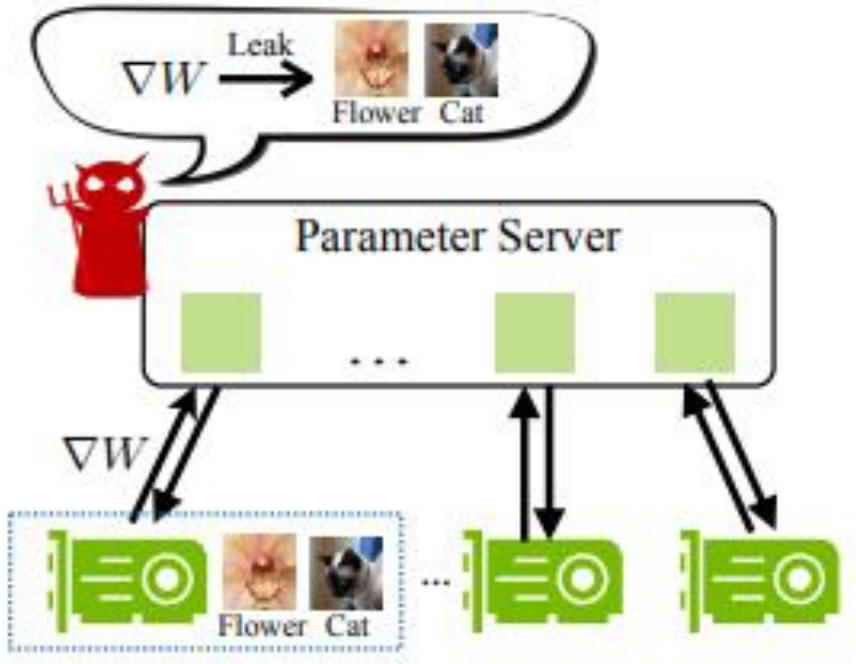
Start with the **initial model weights W_0** shared with all clients.

- Client Training:

Each client trains the model on its **private data X** and computes the **gradients ∇W, ∇b** based on the loss function $L\ (W, X)$.

- Attack Objective:

The attacker aims to **recover the private data X** by exploiting the gradients ∇W, ∇b shared with the server after training.



STATE OF THE ART

- Gradient attacks are effective methods for recovering private training data from gradient vectors. They are primarily categorized:
- Optimization-based attacks.

DLG. [1] proposed an algorithm that takes the random “dummy” gradients and corresponding class labels to process them through the forward and backward passes then minimizing the distance between the dummy gradients and the real gradients.

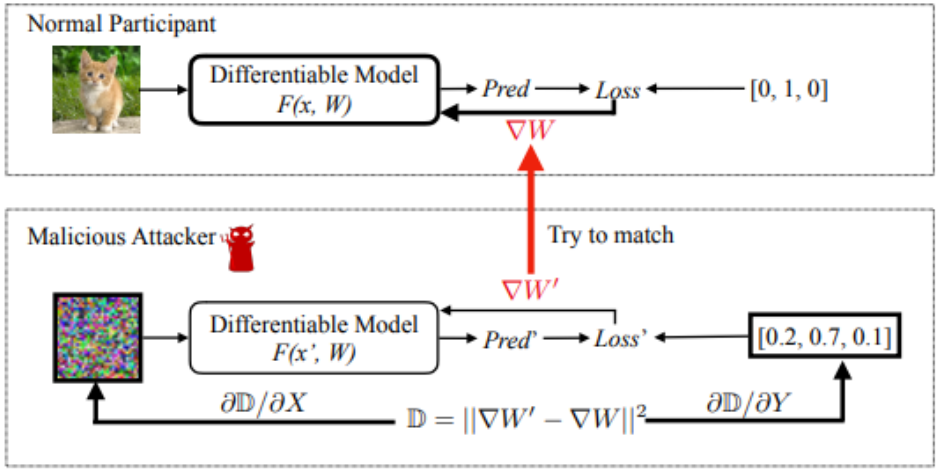
Limitations:

- Only works with specific activation functions.
- Not a deterministic approach, results may vary.
- Heavily influenced by the initialization of the dummy input.

- Analytical-based attacks:

R-GAP [2] introduced an algorithm which that leverages the weights constraints to recover the input data but :

- Limitations:
  - Requires fully invertible activation functions.
  - Underestimate the importance of gradients constrains in the convolution layers.



CONTRIBUTIONS

- Key contributions :
- Introduced R-CONV, an advanced analytical method that overcomes the limitations of R-GAP in data reconstruction.
  - And we demonstrate that the attack is feasible even in the presence of non-fully invertible functions
- Emphasized the importance of gradient constraints, which can be more revealing than weight constraints in certain layers.
- Analyzed how convolutional layer parameters (e.g., kernel size, stride, padding) affect the success of gradient-based attacks.
- Demonstrated that current analytical methods for estimating gradient attack risks lack accuracy, as they underestimate the role of gradient constraints.

METHODOLOGY

- We structure our approach into three key phases:
- Gradient Computation and Input Reconstruction from Fully Connected Layer.
- Propagating Gradient Through Activation Function.
- Gradient Computation and Input Reconstruction in Convolution Layer

METHODOLOGY

- Gradient Computation and Input Reconstruction from Fully Connected Layer :

The output of the layer can be expressed as:

$$Z = WX + b,$$

and the output for a node m can be represented as:

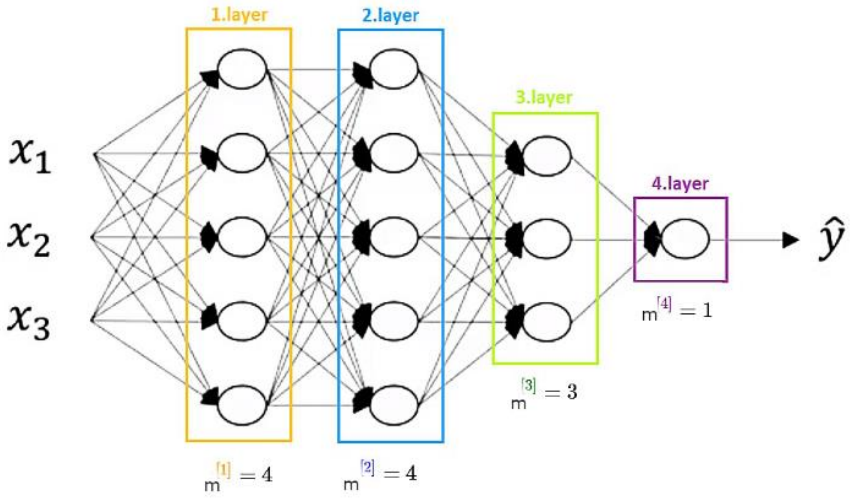
$$z_m = \sum_{i=1}^{i} (x_i w_{im}) + b_m$$

$$\frac{\partial z_m}{\partial b_m} = 1, \frac{\partial z_m}{\partial x_n} = w_{nm} \tag{1}$$

Based on the chain rule we can write:

$$\frac{\partial \ell}{\partial b_m} = \frac{\partial \ell}{\partial z_m} \times \frac{\partial z_m}{\partial b_m} \tag{2}$$

$$\frac{\partial \ell}{\partial x_n} = \frac{\partial \ell}{\partial z_m} \times \frac{\partial z_m}{\partial x_n} \tag{3}$$



METHODOLOGY

- Gradient Computation and Input Reconstruction from Fully Connected Layer:

By substituting from equation 1 into equation 2 we can write,

$$\frac{\partial \ell}{\partial z_m} = \frac{\partial \ell}{\partial b_m} \tag{4}$$

and by substituting from equation (1) and (4) in (3) we can get:

$$\frac{\partial \ell}{\partial x_n} = \frac{\partial \ell}{\partial b_m} \times w_{nm} \tag{5}$$

Equation (5) represents the gradient w.r.t. one node. To complete the total gradient w.r.t. the input x_n can be derived from gradients w.r.t. W and gradients w.r.t. b:

$$\frac{\partial \ell}{\partial x_n} = \sum_{c=1}^{c} \frac{\partial \ell}{\partial x_n} \tag{6}$$

METHODOLOGY

- Gradient Computation and Input Reconstruction from Fully Connected Layer:

$$\frac{\partial \ell}{\partial w_{nm}} = \frac{\partial \ell}{\partial z_m} \times \frac{\partial z_m}{\partial w_{nm}}, \frac{\partial z_m}{\partial w_{nm}} = x_n$$

$$\frac{\partial \ell}{\partial w_{nm}} = \frac{\partial \ell}{\partial z_m} x_n$$

$$x_n = \frac{\partial \ell}{\partial w_{nm}} / \frac{\partial \ell}{\partial z_m}$$

and by substituting from equation (4) one obtains:

$$x_n = \frac{\partial \ell}{\partial w_{nm}} / \frac{\partial \ell}{\partial b_m}$$

RESEARCH ACTIVITY 1 – METHODOLOGY

- Propagating Gradient Through Activation Function:

$$\frac{\partial \ell}{\partial O} = \frac{\partial \ell}{\partial X} \times \frac{\partial X}{\partial O}$$

$$\frac{\partial \ell}{\partial O} = \frac{\partial \ell}{\partial X} \times A'(O)$$

Table 1: Derivative of activation functions.

Name	Equation	Derivative expressing in X
Sigmoid	$A(O) = \frac{1}{1+e^{-O}}$	$A'(O) = X(1-X)$
Tanh	$A(O) = \tanh(O)$	$A'(O) = 1 - X^2$
ArcTan	$A(O) = \tan^{-1}(O)$	$A'(O) = \frac{1}{1-\tan(X)^2}$
SoftPlus	$A(O) = \log_e(1+e^O)$	$A'(O) = \frac{1}{1+\ln(e^X-1)}$
ReLU	$A(O) = \begin{cases} O & \text{if } O > 0 \\ 0 & \text{if } O \leq 0 \end{cases}$	$A'(O) = \begin{cases} 1 & \text{if } X > 0 \\ 0 & \text{if } X = 0 \end{cases}$
Leaky ReLU	$A(O) = \begin{cases} O & \text{if } O \geq 0 \\ 0.01O & \text{if } O < 0 \end{cases}$	$A'(O) = \begin{cases} 1 & \text{if } X \geq 0 \\ 0.01 & \text{if } X < 0 \end{cases}$
Parameteric ReLU	$A(O) = \begin{cases} O & \text{if } O \geq 0 \\ \alpha O & \text{if } O < 0 \end{cases}$	$A'(O) = \begin{cases} 1 & \text{if } X \geq 0 \\ \alpha & \text{if } X < 0 \end{cases}$
ELU	$A(O) = \begin{cases} O & \text{if } O \geq 0 \\ \alpha(e^O - 1) & \text{if } O < 0 \end{cases}$	$A'(O) = \begin{cases} 1 & \text{if } X \geq 0 \\ X + \alpha & \text{if } X < 0 \end{cases}$

RESEARCH ACTIVITY 1 – METHODOLOGY

- Gradient Computation and Input Reconstruction in Convolution Layer

$$
\begin{aligned}
o_{1,1} &= x_1 w_{1,1} + x_2 w_{1,2} + x_4 w_{1,4} + x_5 w_{1,5} \\
o_{1,2} &= x_2 w_{1,1} + x_3 w_{1,2} + x_5 w_{1,4} + x_6 w_{1,5} \\
o_{1,3} &= x_4 w_{1,1} + x_5 w_{1,2} + x_7 w_{1,4} + x_8 w_{1,5} \\
o_{1,4} &= x_5 w_{1,1} + x_6 w_{1,2} + x_8 w_{1,4} + x_9 w_{1,5} \qquad (9)
\end{aligned}
$$

Gradient Computation Applying the chain rule, the gradient w.r.t. x_1 is expressed as:

$$\frac{\partial \ell}{\partial x_1} = \frac{\partial \ell}{\partial o_{1,1}} \times \frac{\partial o_{1,1}}{\partial x_1}$$

And from previous equations we can obtain $\frac{\partial o_{1,1}}{\partial x_1} = w_{1,1}$. Thus,

$$\frac{\partial \ell}{\partial x_1} = \frac{\partial \ell}{\partial o_{1,1}} \times w_{1,1}$$

RESEARCH ACTIVITY 1 – METHODOLOGY

- Gradient Computation and Input Reconstruction in Convolution Layer

Input Reconstruction Applying the chain rule the gradient w.r.t. $w_{1,1}$ is expressed as:

$$\frac{\partial \ell}{\partial w_{1,1}} = \frac{\partial \ell}{\partial o_{1,1}} \times \frac{\partial o_{1,1}}{\partial w_{1,1}} + \frac{\partial \ell}{\partial o_{1,2}} \times \frac{\partial o_{1,2}}{\partial w_{1,1}} + \frac{\partial \ell}{\partial o_{1,3}} \times \frac{\partial o_{1,3}}{\partial w_{1,1}} + \frac{\partial \ell}{\partial o_{1,4}} \times \frac{\partial o_{1,4}}{\partial w_{1,1}}$$

From equation (9) we can obtain:

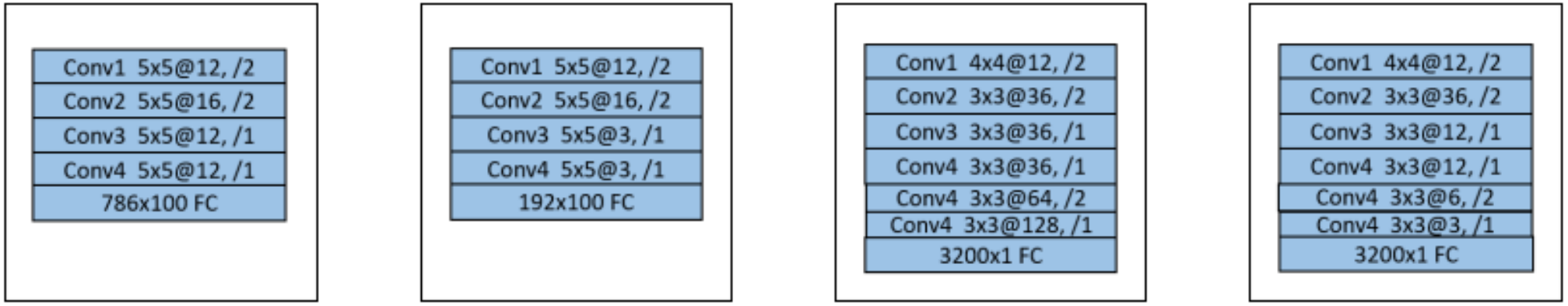
$$\frac{\partial o_{1,1}}{\partial w_{1,1}} = x_1, \frac{\partial o_{1,2}}{\partial w_{1,1}} = x_2, \frac{\partial o_{1,3}}{\partial w_{1,1}} = x_4, \frac{\partial o_{1,1}}{\partial w_{1,1}} = x_5$$

Thus,

$$\frac{\partial \ell}{\partial w_{1,1}} = \frac{\partial \ell}{\partial o_{1,1}} \times x_1 + \frac{\partial \ell}{\partial o_{1,2}} \times x_2 + \frac{\partial \ell}{\partial o_{1,3}} \times x_4 + \frac{\partial \ell}{\partial o_{1,1}} \times x_5$$

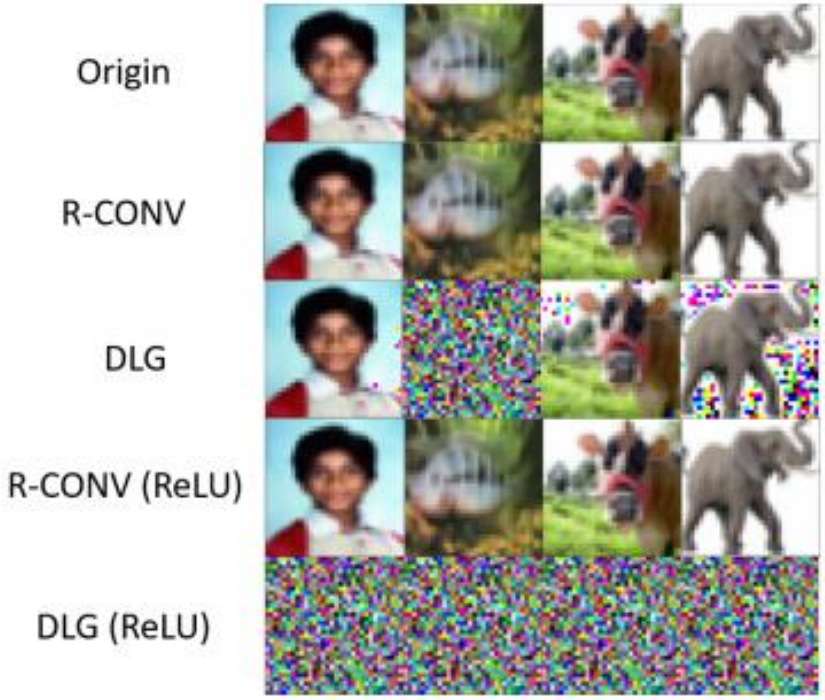
RESEARCH ACTIVITY 1 – RESULTS

- To illustrate the effectiveness of our R-CONV method, we compared it with DLG and R-GAP [16] using three different datasets: CIFAR-100, CIFAR-10, and MNIST. We employed Four different CNN architectures.

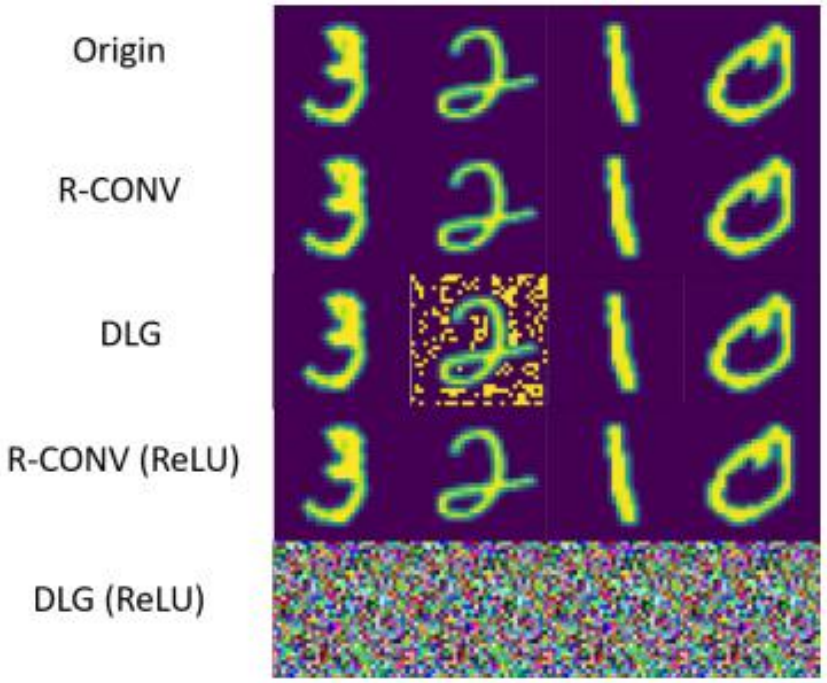


(a) LeNet (b) LeNet-O (c) CNN6 (d) CNN6-O

RESULTS



(a) CIFAR100



(b) MNIST

RESULTS

RESULTS

Table 2: Comparison of the proposed R-CONV method with state-of-the-art analytical (R-GAP) and optimization-based (DLG) methods in terms of average MSE, PSNR, and reconstruction time. *Our proposed method outperforms these state-of-the-art methods in all the considered metrics.*

Method	MSE	PSNR (dB)	Time (s)
Average Computed for Images Depicted in Figure 2.			
R-CONV	$2.2 \times 10^{-7} \pm 3.64 \times 10^{-9}$	114.68 ± 5.5	6.33 ± 2.3
DLG	0.0933 ± 0.05	62.62 ± 8.5	60.66 ± 5.58
Average Computed for Images Depicted in Figure 3.			
R-CONV	$2.88 \times 10^{-9} \pm 2.44 \times 10^{-10}$	150.12 ± 4.5	2.494 ± 1.66
R-GAP	0.0056 ± 0.008	76.73 ± 6.88	232.45 ± 12.44

MITIGATE THE RISK OF GRADIENT ATTACKS

- Suppose the dimension for the input is $(H * H * N)$ the number of filter is F, and the kernel size is K, with the stride S and padding P:

$$|A^l| = (\frac{H + 2P - K}{S} + 1) * F$$

$$|B^l| = K^2 * F$$

- We must ensure that the following conditions hold:

$$|X^l| > |A^l| + |B^l|.$$

FUTURE DIRECTIONS

1. Extend the Attack to Batches and High-Dimensional Images
    - Scale the attack to work on mini-batch processing and high-resolution images to better reflect real-world scenarios.
2. Develop a Robust Defense Mechanism
    - Leverage data leakage sources to design a defense mechanism that can simultaneously counter multiple types of attacks.


Q&A
THANK YOU !