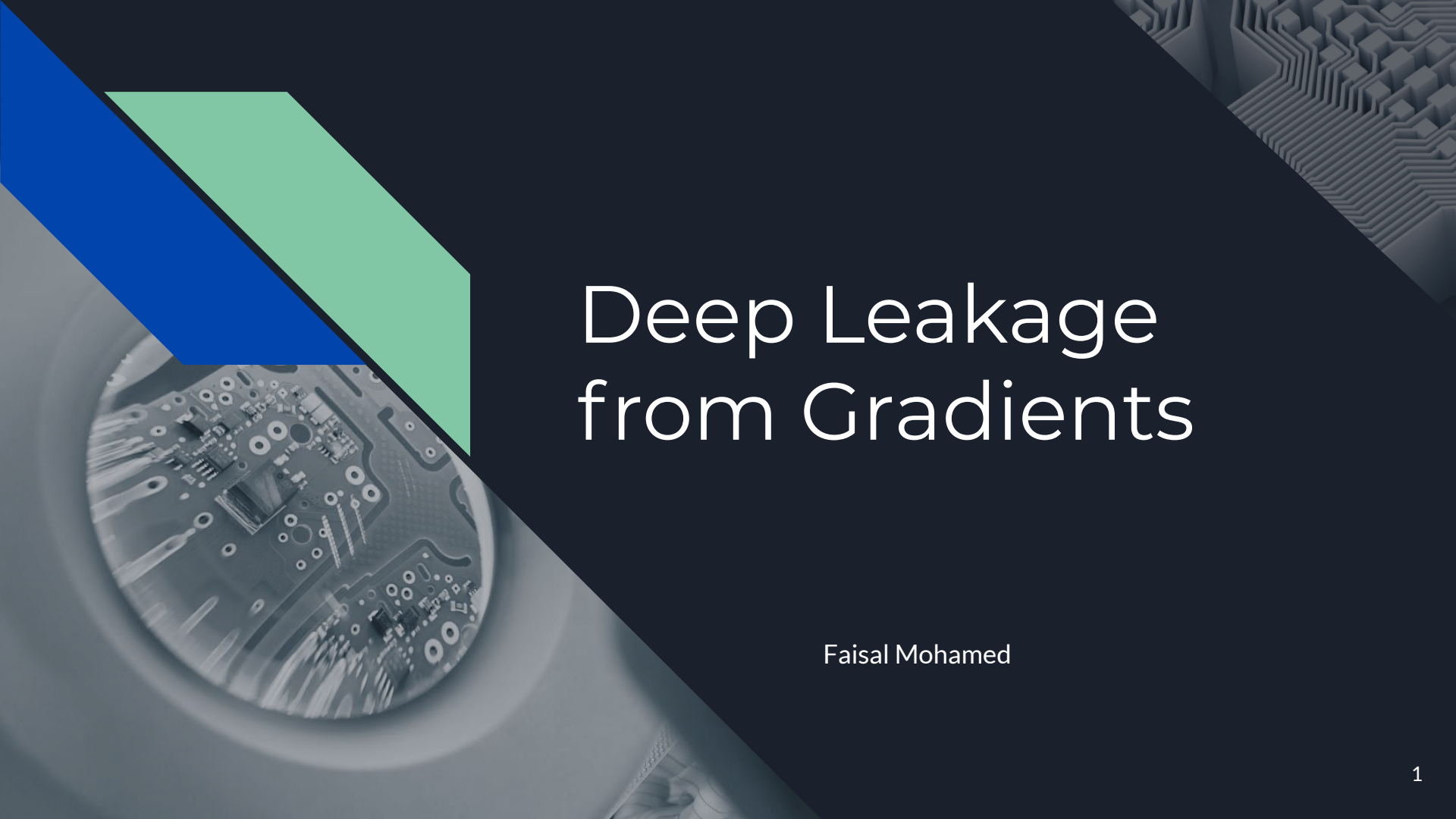


Deep Leakage from Gradients

Faisal Mohamed

Core Idea

Implementation

Results

Core Idea

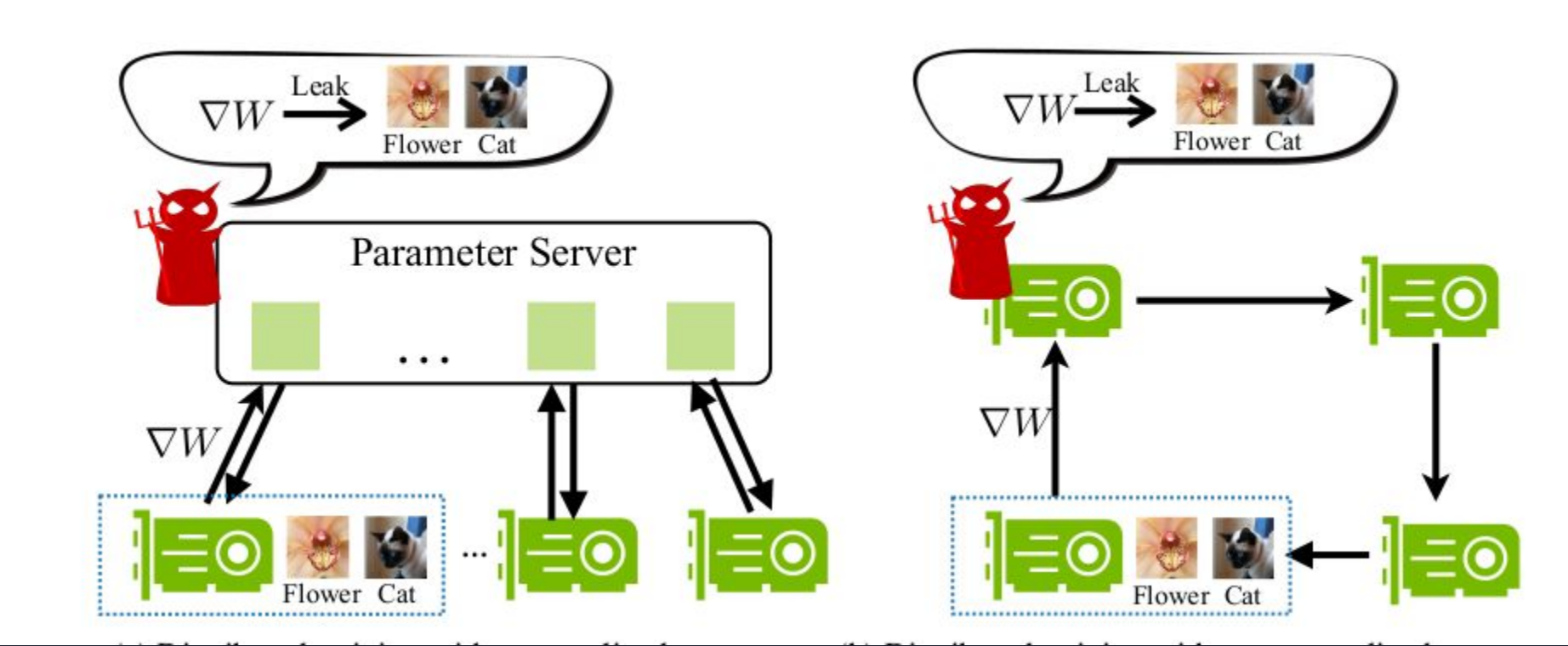
Implementation

Results

Core Idea

- Obtain private training data from publicly shared gradients.
- Raise awareness about the safety of sharing gradients.

Core Idea



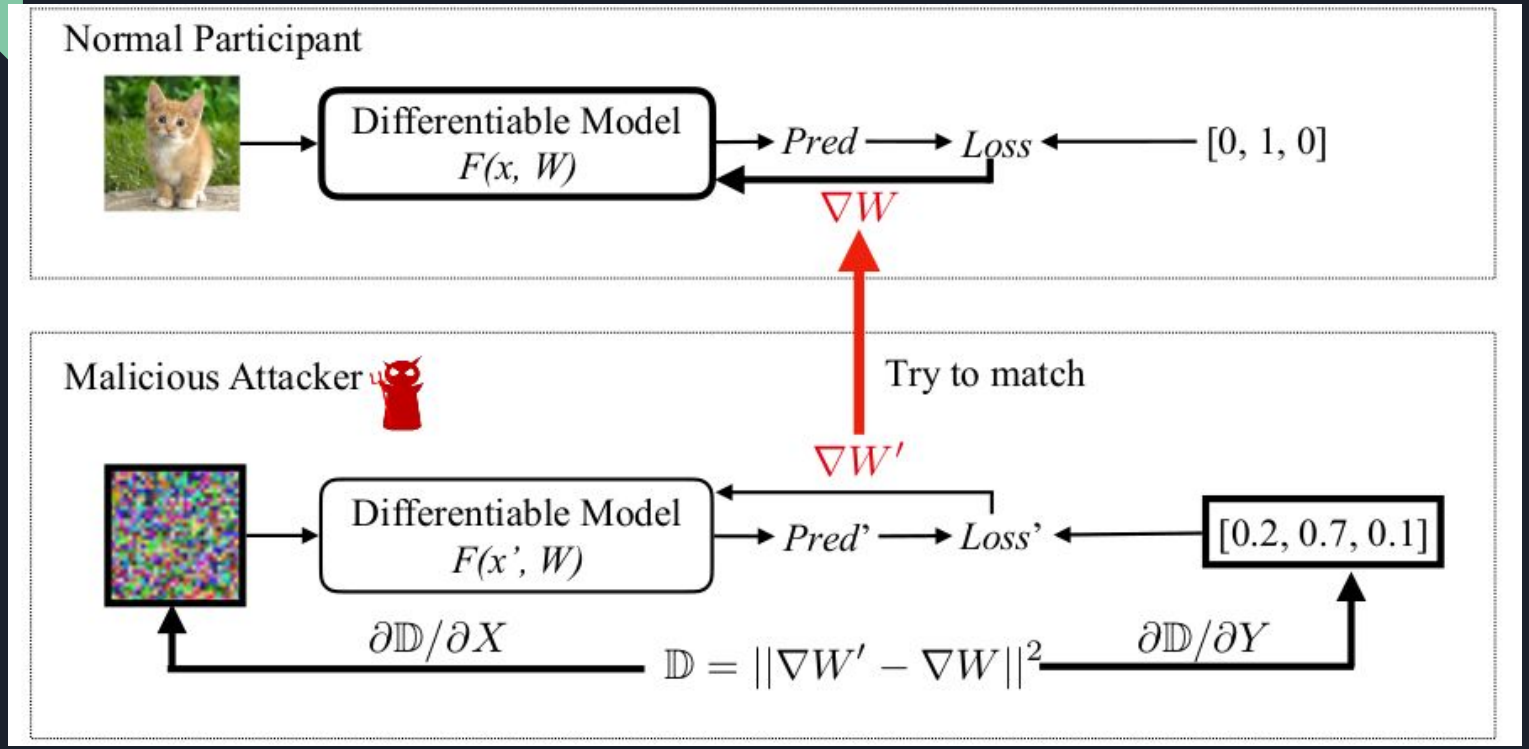
https://arxiv.org/abs/1906.08935

Core Idea

Implementation

Results

Implementation



https://arxiv.org/abs/1906.08935

Implementation

Algorithm 1 Deep Leakage from Gradients.

Input: $F(\mathbf{x}; W)$: Differentiable machine learning model; W: parameter weights; ∇W: gradients calculated by training data

Output: private training data $\mathbf{x}, \mathbf{y}$

procedure DLG(F, W, ∇W)

  $\mathbf{x'}_1 \leftarrow \mathcal{N}(0,1)$, $\mathbf{y'}_1 \leftarrow \mathcal{N}(0,1)$ ▷ Initialize dummy inputs and labels.

  for $i \leftarrow 1$ to n **do**

    $\nabla W'_i \leftarrow \partial \ell(F(\mathbf{x'}_i, W_t), \mathbf{y'}_i)/\partial W_t$ ▷ Compute dummy gradients.

    $\mathbb{D}_i \leftarrow ||\nabla W'_i - \nabla W||^2$

    $\mathbf{x}'_{i+1} \leftarrow \mathbf{x}'_i - \eta \nabla_{\mathbf{x}'_i} \mathbb{D}_i$, $\mathbf{y}'_{i+1} \leftarrow \mathbf{y}'_i - \eta \nabla_{\mathbf{y}'_i} \mathbb{D}_i$ ▷ Update data to match gradients.

  end for

  return $\mathbf{x}'_{n+1}, \mathbf{y}'_{n+1}$

end procedure

https://arxiv.org/abs/1906.08935

Core Idea

Implementation

Results

Results


0
25
0
50
100
150
200
250
300
0
25
0
50
100
150
200
250
300
350
0
25
0
50
100
150
200
250
300
350