DAPter: Preventing User Data Abuse in Deep Learning Inference Services

Hao Wu¹, Xuejin Tian¹, Yuhang Gong¹, Xing Su¹, Minghao Li¹², Fengyuan Xu¹*

¹Nanjing University  ²Cornell University

*Corresponding author: fengyuan.xu@nju.edu.cn
Deep Learning Inference Service (DLIS) prospers.
Deep Learning Inference Service (DLIS) prospers.
DLIS Scenario
DLIS Scenario
DLIS Scenario
Data abuse issue
Data abuse issue

Abusability!!!

Text | Audio | Video | Image

DLIS
Data abuse issue

Data abuse is about the rights of data owners in the context of DLIS.

1. Infer private info.
2. Train new models.
Problem Requirements
Problem Requirements

**Honest Provider:**
- Attract more customers
- Reduce potential risks of violating laws (GDPR, CCPA)

![Diagram showing user data and results](image)
Problem Requirements

Honest Provider:
- Attract more customers
- Reduce potential risks of violating laws (GDPR, CCPA)

S1. Not visually recognizable
S2. Only retain necessary features
S3. Can’t be reversed
Problem Requirements

Honest Provider:
- Attract more customers
- Reduce potential risks of violating laws (GDPR, CCPA)

S1. Not visually recognizable
S2. Only retain necessary features
S3. Can’t be reversed

DLIS

U1. Maintain Accuracy
U2. No Changes
U3. Efficient

- Text
- Audio
- Video
- Image

User Data

Results

S1.
S2.
S3.
**Problem Requirements**

**Honest Provider:**
- Attract more customers
- Reduce potential risks of violating laws (GDPR, CCPA)

**Balance**
- Security
  - Weak security solution
    - DP, MP, PAN
- Usability
  - Low usability solution
    - TEE, FHE
Our solution DAPter
Our solution DAPter

DAPter: Preventing User Data Abuse in Deep Learning Inference Services· WWW 2021

DAPter: A lightweight DLIS-input converter at the end user side.
Our solution DAPter

Abuse-prevented !!!

DAPter

A lightweight DLIS-input converter at the end user side.
DAPter Use Case
DAPter Use Case

Abusable

Before Protection

DLIS

User Data

Results

After Protection

Abuse-prevented

DLIS

User Data

Results

DAPter

A lightweight DLIS-input converter at the end user side.
Workflow

A user-side entropy reduction approach to prune information not relevant to the target DLIS in user data.
Workflow

A user-side entropy reduction approach to prune information not relevant to the target DLIS in user data.
Workflow

A user-side *entropy reduction* approach to *prune information* not relevant to the target DLIS in user data.
A user-side entropy reduction approach to prune information not relevant to the target DLIS in user data.
A user-side entropy reduction approach to prune information not relevant to the target DLIS in user data.

**Workflow**

**S1.** Not visually recognizable  
**S2.** Only retain necessary features  
**S3.** Can’t be reversed
Workflow

A user-side entropy reduction approach to prune information not relevant to the target DLIS in user data.
A user-side entropy reduction approach to prune information not relevant to the target DLIS in user data.

1. A lightweight generative model
2. A data abuse prevention loss

Key Design

S1. Not visually recognizable
S2. Only retain necessary features
S3. Can’t be reversed

U1. Maintain Accuracy
U2. No Changes
U3. Efficient
Training Structure & Model Architecture

Original Image → DAPter → Converted Image → Target DLIS → Result

Ground Truth (e.g., male)

(e.g., gender inference)
Training Structure & Model Architecture

Ground Truth (e.g., male)

Original Image

DAPter

Converted Image

Target DLIS (e.g., gender inference)

Result (e.g., male)

$L_{acc}$

Inference accuracy loss

Entropy reduction loss

$L_{\eta}$
Training Structure & Model Architecture

Original Image → DAPter → Converted Image → Target DLIS (Frozen) → Result (e.g., male)

- Inference accuracy loss: $L_{acc}$
- Entropy reduction loss: $L_{\eta}$
- For balance: $L_{dap}$

Ground Truth (e.g., male)
Training Structure & Model Architecture

Inference accuracy loss
Entropy reduction loss
For balance

\[ L_{\text{dap}} = \lambda \times L_{\eta} + (1 - \lambda) \times L_{\text{acc}}, \lambda \in (0, 1) \]
**Training Structure & Model Architecture**

1. A symmetrical U-Net like architecture.
2. Input and output are of the same size.
3. "Copy" connection captures the high-level semantic info and low-level spatial info.

**Model Architecture**

- **Ground Truth** (e.g., male)
- **Original Image**
- **DAPter**
- **Converted Image**
- **Target DLIS** (e.g., gender inference)
- **Result** (e.g., male)

**Inference Accuracy Loss**

\[ L_{\text{acc}} \]

**Entropy Reduction Loss**

\[ L_\eta \]

**Balance Loss**

\[ L_{\text{dap}} = \lambda * L_\eta + (1 - \lambda) * L_{\text{acc}}, \lambda \in (0, 1) \]
Data Abuse Prevention Loss

Minimize the piece of pixel-wise entropy that contributes little to the high-level features.

\[ L_{dap} = \lambda * L_\eta + (1 - \lambda) * L_{acc}, \lambda \in (0, 1) \]
Data Abuse Prevention Loss

Minimize the piece of pixel-wise entropy that contributes little to the high-level features.

\[ L_{dap} = \lambda \cdot L_{\eta} + (1 - \lambda) \cdot L_{acc}, \lambda \in (0, 1) \]

\( L_{acc} \) measures the inference accuracy of the target DLIS.

\( L_{\eta} \) measures the pixel-wise entropy \( (H_I = -\sum_i p_i \log p_i) \) in input data. \( p_i \) is the occurrence possibility of \( i \).
Data Abuse Prevention Loss

Minimize the piece of pixel-wise entropy that contributes little to the high-level features.

\[
L_{dap} = \lambda * L_\eta + (1 - \lambda) * L_{acc}, \lambda \in (0, 1)
\]

**Proof can be found in our paper**

\(L_{acc}\) measures the inference accuracy of the target DLIS.

\(L_\eta\) measures the pixel-wise entropy \((H_I = -\sum_i p_i \log p_i)\) in input data. \(p_i\) is the occurrence possibility of \(i\).

**Support statement**. By enlarging the occurrence possibility of a specific pixel value, the upper bound of an image’s pixel-wise entropy can be reduced.
Data Abuse Prevention Loss

Minimize the piece of pixel-wise entropy that contributes little to the high-level features.

\[ L_{dap} = \lambda \ast L_\eta + (1 - \lambda) \ast L_{acc}, \lambda \in (0, 1) \]

- \( L_{acc} \) measures the inference accuracy of the target DLIS.
- \( L_\eta \) measures the pixel-wise entropy \( (H_I = -\sum_i p_i \log p_i) \) in input data. \( p_i \) is the occurrence possibility of \( i \).

Support statement*. By enlarging the occurrence possibility of a specific pixel value, the upper bound of an image’s pixel-wise entropy can be reduced.

\[ L_\eta = \sum_I \eta(I, I_{ref}) \]

- \( \eta \) is L1 norm; \( I \) is the converted image; \( I_{ref} \) is the reference image with each pixel equaling to (R128, G128, B128).

*Proof can be found in our paper
**Hyperparameter $\lambda$ Exploration**

$$L_{dap} = \lambda * L_\eta + (1 - \lambda) * L_{acc}, \lambda \in (0, 1)$$

A larger $\lambda$ lets DAPter remove more entropy but leads to a low DLIS accuracy.
**Hyperparameter $\lambda$ Exploration**

$$L_{dap} = \lambda \cdot L_\eta + (1 - \lambda) \cdot L_{acc}, \lambda \in (0, 1)$$

A larger $\lambda$ lets DAPter remove more entropy but leads to a low DLIS accuracy.

(a) Relationship between $\lambda$ and Accuracy

(b) Relationship between $\lambda$ and Entropy
Heyperparameter $\lambda$ Exploration

$$L_{dap} = \lambda \ast L_\eta + (1 - \lambda) \ast L_{acc}, \lambda \in (0, 1)$$

A larger $\lambda$ lets DAPter remove more entropy but leads to a low DLIS accuracy.

$\lambda = 0.9$ is a sweet point to balance security and usability.
Conversion Quality

To show that DAPter can remove the unnecessary features and retain the useful features, we generate saliency map (SM) to measure which part of the input supports the DLIS through Grad-CAM.
Conversion Quality

To show that DAPter can remove the unnecessary features and retain the useful features, we generate saliency map (SM) to measure which part of the input supports the DLIS through Grad-CAM.

Results are visualize below. From left to right is original image, sm of DLIS, protected image, sm of DAPter-enabled DLIS.

(a) Arched Eyebrow Inference
(b) Wearing Glasses Inference
(c) Gender Inference
Security - Auto Recognition Attack

The adversary can use SOTA DL model to label the entropy-reduced outputs of DAPter.
Security - Auto Recognition Attack

The adversary can use SOTA DL model to label the entropy-reduced outputs of DAPter.

**Ground Truth**
(e.g., black hair)

**Original Image**
(e.g., trained for gender inference)

**DAPter**

**Converted Image**
(e.g., hair color inference)

**Adv. Model**
(e.g., hair color inference)

**Result**
(e.g., unknown)

**Attack Accuracy**

Case 1: Attack tasks have no correlation with the targeted task.

Case 2: Attack tasks have correlations with the targeted task.

Accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cifar20:100</th>
<th>LFW</th>
<th>ImageNet8:32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Task</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attack Task</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy

<table>
<thead>
<tr>
<th>(1) Bangs</th>
<th>(2) Blond Hair</th>
<th>(3) Receding Hairline</th>
<th>(4) Goatee</th>
<th>(5) Big Nose</th>
<th>(6) No Beard</th>
<th>(7) Heavy Makeup</th>
<th>(8) Wearing Lipstick</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc_ideal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc_ideal__</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAPter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Security - Image Reconstruction Attack

The adversary can use SOTA DL model to reconstruct the original image from the protected one.
Security - Image Reconstruction Attack

The adversary can use SOTA DL model to reconstruct the original image from the protected one.

(a) Chubby Inference Task
(b) Wearing Glasses Inference Task
(c) Wearing Lipstick Inference Task
Usability Evaluation

Backend Throughput:
- Compare to TEE-based solution: \(2.5x\sim 50x\),
- Compare to FHE-based solution: \(1000x\).

Bandwidth Usage:
- \(2.1x\sim 41x\) better (measured with LFW, ImageNet, CelebA, Cifar10).

Latency Overhead:
- \(109\text{ms}\) (Snapdragon 855 Plus), \(292\text{ms}\) (Kirin 960), and \(309\text{ms}\) (Helio X30).

No DLIS backend change is needed!!!
Take away

First investigate the data abuse issue in the scenario of DLIS.
Take away

First investigate the data abuse issue in the scenario of DLIS.

A user-side entropy reduction approach to prevent data abuse in DLIS context.
Take away

**First** investigate the data abuse issue in the scenario of DLIS.

A user-side entropy reduction approach to **prevent data abuse** in DLIS context.

- **Security**
  - S1. Not visually recognizable
  - S2. Only retain necessary features
  - S3. Can't be reversed

- **Usability**
  - U1. Maintain Accuracy
  - U2. No Changes
  - U3. Efficient

A lightweight user-side add-on.

DAPter: Preventing User Data Abuse in Deep Learning Inference Services· WWW 2021
Take away

First investigate the data abuse issue in the scenario of DLIS. A user-side entropy reduction approach to prevent data abuse in DLIS context.

Security

S1. Not visually recognizable
S2. Only retain necessary features
S3. Can’t be reversed

Usability

U1. Maintain Accuracy
U2. No Changes
U3. Efficient

Balance

DAPter: A lightweight user-side add-on.

Results

User Data

Thank you for attention!