How Does a Deep Learning Model Architecture Impact Its Privacy?

— A Comprehensive Study of Privacy Attacks on CNNs and Transformers

Ming Ding {ming.ding@data61.csiro.au}
Principal Research Scientist & Science Lead
Privacy Technology Group, Data61, CSIRO, Australia
Aug 2024

Australia's National Science Agency

Guangsheng Zhang[1], Bo Liu[1], Huan Tian[1], Tianqing Zhu[1], Ming Ding[2], Wanlei Zhou[3]

Agenda

- Background and Motivation
- Methodology and Key Findings
- Appendix: More Experimental Results

CSIRO

Global AI Regulation Activities

Apr 2021

Topics
European Parliament

EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's

- Risk-based approach
- Safe, transparent, traceable, non-discriminatory and environmentally friendly

Oct 2023

WHITE HOUSE

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

- Establish standards for AI safety and security
- Protects Americans' privacy
- Advances equity and civil rights, etc

Dec 2023

News
European Parliament

Artificial Intelligence Act: MEPs adopt landmark law

- The EU Parliament and Council reached an agreement on the Act
- The expected year of implementation is set to 2026

Jan 2024

Supporting responsible AI: discussion paper

- DISR gave its interim response to the consultation for Safe and Responsible AI
- Adopt a risk-based framework

Privacy Attacks – Membership Inference Attack

Original data *x*
(sensitive info: *s*
non-sensitive info: *u*)

→ AI/ML function →

Output/Observation *y*
(data, query answer, AI/ML models/parameters, etc.)

- Membership Inference Attack (MIA): $\Pr[h \in x_{\text{train}} | y]$
- y: Victim model's prediction results or confidence scores
- Method: NN based (Shadow training, prediction confidence scores), Likelihood based (LiRA, Likelihood ratio attack, multi shadow models)

Privacy Attacks – Attribute Inference Attack

Original data *x*
(sensitive info: *s*
non-sensitive info: *u*)

$\Rightarrow$ AI/ML function $\Rightarrow$

Output/Observation *y*
(data, query answer, AI/ML
models/parameters, etc.)

- Attribute Inference Attack (AIA): $I(s; y)$
- y: Victim model's intermediate features
- Method: Train an attack model based on features

Privacy Attacks – Gradient Inversion Attack

Original data *x*
(sensitive info: *s*
non-sensitive info: *u*)

$\Rightarrow$ AI/ML function $\Rightarrow$

Output/Observation *y*
(data, query answer, AI/ML models/parameters, etc.)

- Gradient Inversion Attack (GIA): $\Pr[x|y]$
- y: Victim model's gradients
- Method: Optimization between gradients and reconstructed samples

Motivation of Our Work

- Privacy attack performance varies from model to model, which cannot be solely explained by model's overfitting level.

- Does the design of a model's **architecture** play a role in its **privacy weaknesses**?

Agenda

- Background and Motivation
- Methodology and Key Findings
- Appendix: More Experimental Results

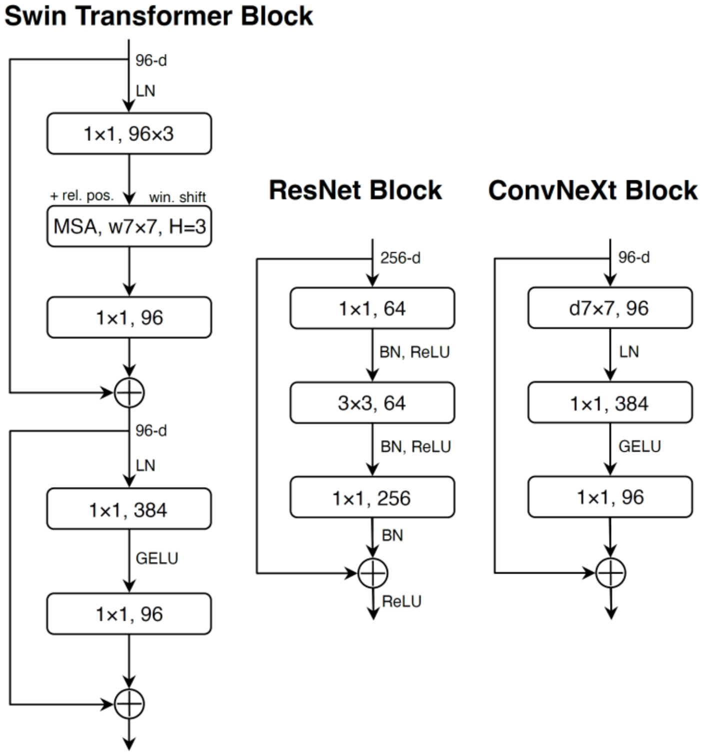
Our Methodology

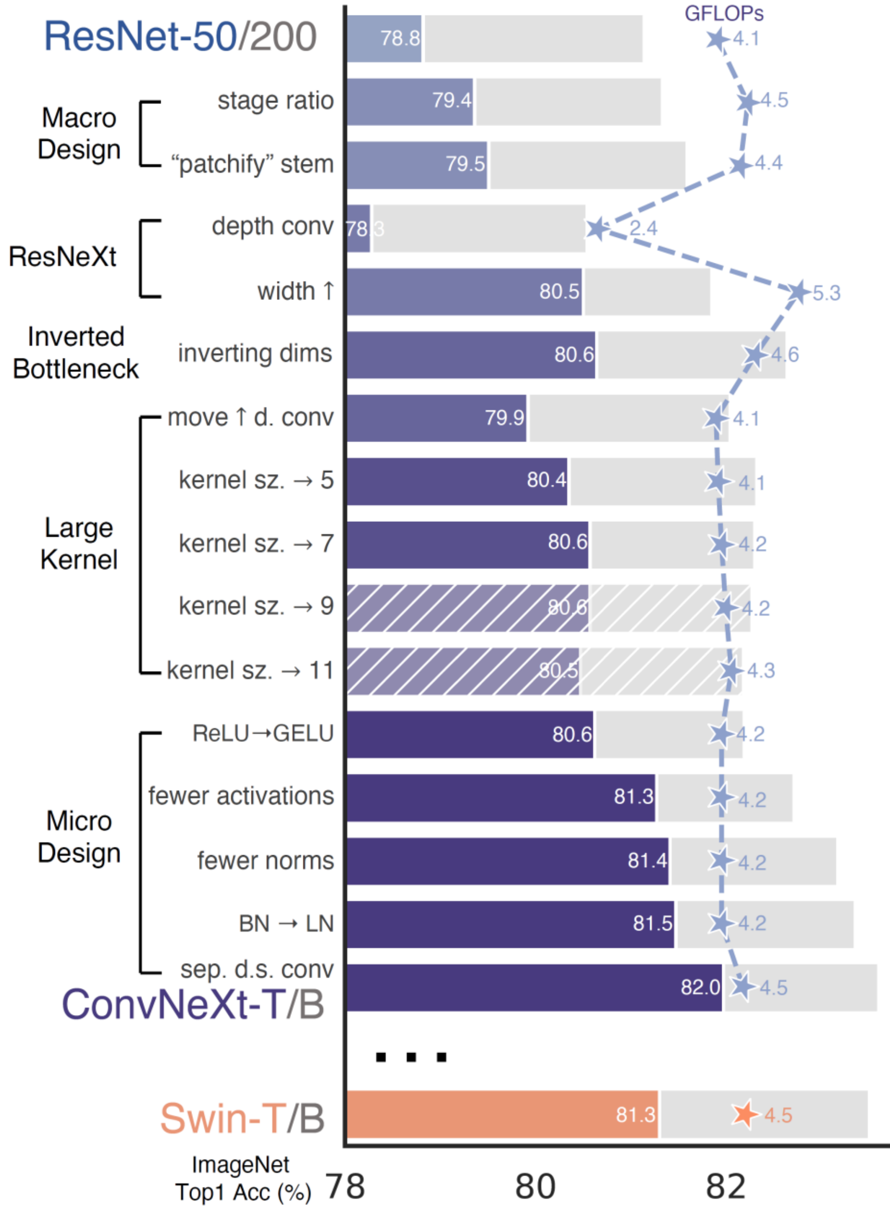
- **Part I**: Conduct a head-to-head comparison of CNNs and transformers
  - Victim CNNs: ResNet-50, ResNet-101
  - Victim Transformers: Swin-T, Swin-S
  - Three privacy attack methods: MIA, AIA, and GIA
  - **Fair comparison: Comparable model sizes, over-fitting levels, primary-task accuracy**

- **Part II**: Morph a CNN to a transformer-like network **step by step**, and identify the steps that introduce significant privacy risks

Morph ResNet-50 to ConvNeXt-T

- Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A convnet for the 2020s." CVPR 2022.





Three Key Features of ResNets/CNNs

1. **Convolution** = Cross-correlation

“Attention” in transformers

- Convolution: $f(t) * g(t) = \int_{-\infty}^{+\infty} f(\tau) g(t - \tau) d\tau$
- Cross-correlation: $f(t) \star g(t) = \int_{-\infty}^{+\infty} f(\tau) g(t + \tau) d\tau$
- np.convolve([1, 3, 1, 2, 3, 3, 5, 1, 3], [1, 0, 2]) ➔ [1, 3, 3, 8, 5, 7, 11, 7, 13, 2, 6]
- np.correlate([1, 3, 1, 2, 3, 3, 5, 1, 3], [2, 0, 1], 'full’) ➔ [1, 3, 3, 8, 5, 7, 11, 7, 13, 2, 6]

2. **Residual connections** to mitigate gradient vanishing

“Skip connections” in transformers

3. **1x1 convolution blocks** for dimension reduction or restoration

“Matrices W^Q, W^K, W^O” in transformers

Experimental Settings

- **Datasets**: CIFAR10, CIFAR100, ImageNet1K, CelebA
- **Attack Models**: MLP models for MIA and AIA. For GIA, we optimize input and generate gradients to reconstruct the underlying data.
- **Metrics for privacy attacks**:
  - MIA: Attack accuracy, Area under the ROC curve (AUC), etc.
  - AIA: Attack accuracy, macro-F1 score, etc.
  - GIA: Mean squared error (MSE), Peak signal-to-noise ratio (PSNR), Learned perceptual image patch similarity (LPIPS), Structural similarity index measure (SSIM), etc.

Experimental Settings

- **Utility metric** for primary classification task: Task Accuracy
- **Over-fitting level metric** for primary classification task: The accuracy difference between the training and testing of a victim model.
- We conducted **~1.5k** experiments/training instances with ~**1.2k** training hours

Main Findings

- **Transformers exhibit higher vulnerabilities to these privacy attacks than CNNs.**

- Primary causes: Fewer activation layers, the "Patchify" method in the stem layers, and layer-normalization layers make transformers more susceptible to privacy attacks than CNNs.

GIA on 14 Intermediate Models from ResNet-50 to ConvNeXt-T (CIFAR10)

	Utility of the model	Efficacy of the GIA			
Steps	Task acc ↑	MSE ↓	PSNR ↑	LPIPS ↓	**SSIM** ↑
1. ResNet-50					
2. Channel dim					
3. Stage ratio					
4. Patchify					
5. ResNeXtify					
6. Inv bottleneck					
7. Kernel sizes					
8. New stem					
9. ReLU to GELU					
10. Removing Act					
11. Removing BN					
12. BN to LN					
13. Sep downsamp					
14. ConvNeXt					

GIA on 14 Intermediate Models from ResNet-50 to ConvNeXt-T (CIFAR10)

Steps	Task acc ↑	MSE ↓	PSNR ↑	LPIPS ↓	SSIM ↑
	$\underbrace{}$ Utility of the model	Efficacy of the GIA			
1. ResNet-50	0.8220 ± 0.0039	1.5096 ± 0.5538	10.58 ± 1.87	0.1624 ± 0.0613	0.0896 ± 0.0544
2. Channel dim	0.8240 ± 0.0072	1.4706 ± 0.5710	10.74 ± 1.97	0.1724 ± 0.0616	0.0826 ± 0.0405
3. Stage ratio	0.8282 ± 0.0040	1.5286 ± 0.5246	10.56 ± 2.05	0.1834 ± 0.0581	0.0731 ± 0.0613
4. Patchify	0.8293 ± 0.0061	**0.9011 ± 0.4376**	**12.97 ± 2.10**	**0.0867 ± 0.0436**	**0.1727 ± 0.0794**
5. ResNeXtify	0.8397 ± 0.0033	1.2415 ± 0.6934	11.86 ± 2.77	0.1066 ± 0.0391	0.1334 ± 0.0950
6. Inv bottleneck	0.8407 ± 0.0058	1.1123 ± 0.4994	12.06 ± 2.19	0.0989 ± 0.0290	0.1429 ± 0.0844
7. Kernel sizes	0.8432 ± 0.0052	0.8206 ± 0.3543	13.40 ± 2.30	0.0821 ± 0.0355	0.2353 ± 0.0766
8. New stem	0.8459 ± 0.0043	0.5684 ± 0.3564	15.43 ± 3.01	0.0752 ± 0.0381	0.4924 ± 0.1205
9. ReLU to GELU	0.8436 ± 0.0027	1.0540 ± 0.5075	12.42 ± 2.61	0.2422 ± 0.0904	0.1746 ± 0.1166
10. Removing Act	0.8480 ± 0.0064	**0.0215 ± 0.0150**	**29.93 ± 3.58**	**0.0049 ± 0.0026**	**0.9562 ± 0.0224**
11. Removing BN	0.8491 ± 0.0059	0.0198 ± 0.0139	30.57 ± 4.12	0.0045 ± 0.0032	0.9605 ± 0.0232
12. BN to LN	0.8501 ± 0.0031	**0.0049 ± 0.0044**	**36.86 ± 3.96**	**0.0005 ± 0.0003**	**0.9927 ± 0.0064**
13. Sep downsamp	0.8553 ± 0.0070	0.0121 ± 0.0171	33.79 ± 4.69	0.0011 ± 0.0008	0.9859 ± 0.0151
14. ConvNeXt	0.8523 ± 0.0064	0.0177 ± 0.0171	31.88 ± 5.04	0.0032 ± 0.0055	0.9666 ± 0.0451

Intuitions

- Fewer activation layers allow transformers to preserve more information learned from the training data (**non-linear function, hard to reverse-engineer**)
- The "Patchify" method in the stem layers is a non-overlapping convolution process (stride=filter width) that can easily learn information from input data, improving the adversary's attack performance LN layers

Goel, Surbhi, Adam Klivans, and Raghu Meka. "Learning one convolutional layer with overlapping patches." *ICML 2018*.

- Parameters in the LN layers increase the risk of overfitting in the model, potentially exposing sensitive information during privacy attacks

Xu, Jingjing, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. "Understanding and improving layer normalization." *NIPS 2019*.

Conclusion

- We discover that Transformers tend to be more vulnerable to privacy attacks than CNNs.
- We found several primary causes in the transformer model designs that lead to the privacy degradation.
- Privacy protection measures: Insert more activation layers and introduce additional noise to the “privacy-leakage” layers.

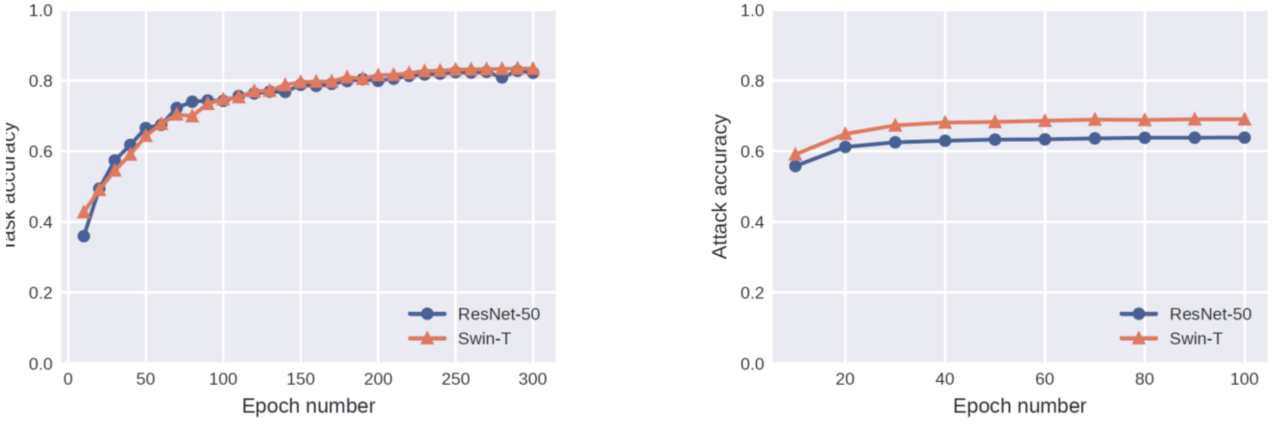
Appendix

Results of MIA

	CIFAR10		CIFAR100	
	Task acc ↑	Attack acc ↑	Task acc ↑	Attack acc ↑
ResNet-50	0.8220 ± 0.0023	0.6385 ± 0.0078	0.5288 ± 0.0083	0.8735 ± 0.0029
Swin-T	0.8335 ± 0.0042	0.6904 ± 0.0052	0.5632 ± 0.0056	0.9340 ± 0.0030
ResNet-101	0.8301 ± 0.0037	0.6317 ± 0.0063	0.5313 ± 0.0074	0.8607 ± 0.0034
Swin-S	0.8258 ± 0.0039	0.6405 ± 0.0075	0.5665 ± 0.0059	0.9357 ± 0.0039

Transform is more vulnerable than CNN when facing privacy attacks

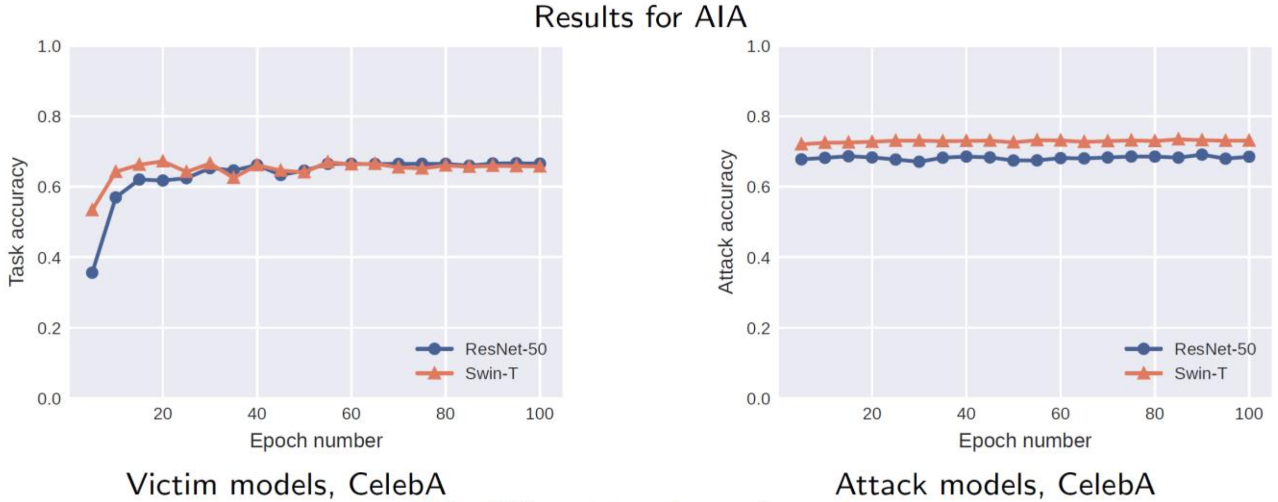
Results for NN-MIA



Victim models, CIFAR10 Attack models, CIFAR10
NN-MIA, different numbers of epochs

Results of AIA

	Task acc ↑	Attack acc ↑	Macro F1 ↑
ResNet-50	0.6666 ± 0.0020	0.6854 ± 0.0015	0.3753 ± 0.0012
Swin-T	0.6587 ± 0.0023	0.7312 ± 0.0014	0.5530 ± 0.0019
ResNet-101	0.6431 ± 0.0029	0.6291 ± 0.0023	0.4262 ± 0.0009
Swin-S	0.6569 ± 0.0024	0.7369 ± 0.0036	0.5536 ± 0.0015



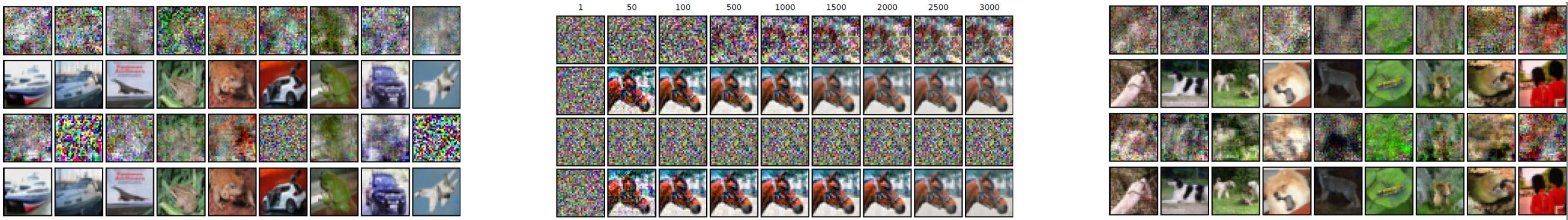
AIA, different numbers of epochs

Transform is more vulnerable than CNN when facing privacy attacks

Results of GIA

	MSE ↓	PSNR ↑	LPIPS ↓	SSIM ↑
ResNet-50	1.3308 ± 0.6507	11.30 ± 2.24	0.1143 ± 0.0403	0.0946 ± 0.0989
Swin-T	0.0069 ± 0.0071	36.24 ± 5.21	0.0012 ± 0.0016	0.9892 ± 0.0118
ResNet-101	1.2557 ± 0.6829	11.58 ± 2.16	0.1461 ± 0.1012	0.0784 ± 0.0675
Swin-S	0.0063 ± 0.0083	37.85 ± 6.15	0.0016 ± 0.0028	0.9878 ± 0.0128

Transform is more vulnerable than CNN when facing privacy attacks



CIFAR10, 3000 iterations   CIFAR10, different iteration numbers   ImageNet1K, 3000 iterations

GIA performance. From the top row to the bottom: ResNet-50, Swin-T, ResNet-101, and Swin-S.

Attack Performance GIA Based on Partial Gradients

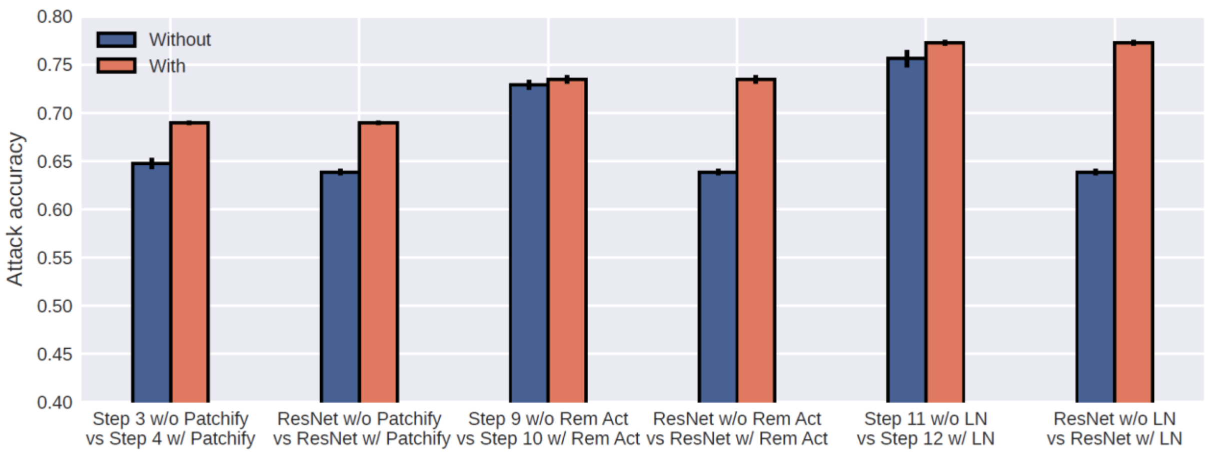
Table 6: The performance of gradient inversion attacks when segmenting ViT-B to make a selection of gradients.

Layers	Num of layers	Params	MSE ↓	PSNR ↑
All	152	85.65M	0.0007 ± 0.0003	43.70 ± 1.84
Stem	4	0.59M	0.0000 ± 0.0000	67.43 ± 5.03
Attention	48	28.34M	0.0020 ± 0.0009	39.61 ± 2.76
MLP	48	56.66M	0.0036 ± 0.0016	36.98 ± 2.59
Norm	48	0.05M	0.0040 ± 0.0018	36.57 ± 2.56
Head	4	0.01M	0.2776 ± 0.2312	19.01 ± 3.89

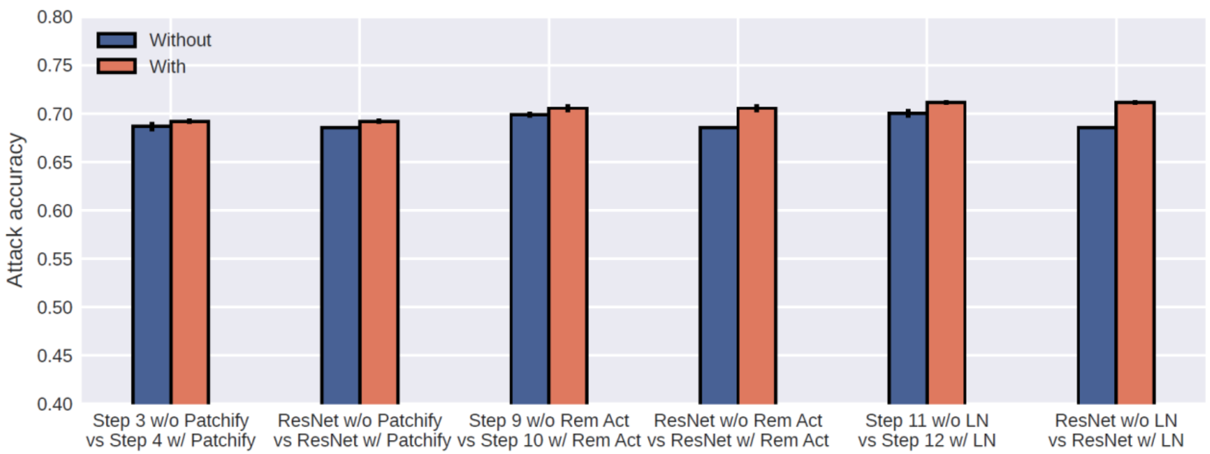
GIA on 14 Intermediate Models from ResNet-50 to ConvNeXt-T

Steps	Task acc ↑	MSE ↓	PSNR ↑	LPIPS ↓	SSIM ↑
1. ResNet-50	0.8220 ± 0.0039	1.5096 ± 0.5538	10.58 ± 1.87	0.1624 ± 0.0613	0.0896 ± 0.0544
2. Channel dim	0.8240 ± 0.0072	1.4706 ± 0.5710	10.74 ± 1.97	0.1724 ± 0.0616	0.0826 ± 0.0405
3. Stage ratio	0.8282 ± 0.0040	1.5286 ± 0.5246	10.56 ± 2.05	0.1834 ± 0.0581	0.0731 ± 0.0613
4. Patchify	0.8293 ± 0.0061	$\mathbf{0.9011 \pm 0.4376}$	$\mathbf{12.97 \pm 2.10}$	$\mathbf{0.0867 \pm 0.0436}$	$\mathbf{0.1727 \pm 0.0794}$
5. ResNeXtify	0.8397 ± 0.0033	1.2415 ± 0.6934	11.86 ± 2.77	0.1066 ± 0.0391	0.1334 ± 0.0950
6. Inv bottleneck	0.8407 ± 0.0058	1.1123 ± 0.4994	12.06 ± 2.19	0.0989 ± 0.0290	0.1429 ± 0.0844
7. Kernel sizes	0.8432 ± 0.0052	0.8206 ± 0.3543	13.40 ± 2.30	0.0821 ± 0.0355	0.2353 ± 0.0766
8. New stem	0.8459 ± 0.0043	0.5684 ± 0.3564	15.43 ± 3.01	0.0752 ± 0.0381	0.4924 ± 0.1205
9. ReLU to GELU	0.8436 ± 0.0027	1.0540 ± 0.5075	12.42 ± 2.61	0.2422 ± 0.0904	0.1746 ± 0.1166
10. Removing Act	0.8480 ± 0.0064	$\mathbf{0.0215 \pm 0.0150}$	$\mathbf{29.93 \pm 3.58}$	$\mathbf{0.0049 \pm 0.0026}$	$\mathbf{0.9562 \pm 0.0224}$
11. Removing BN	0.8491 ± 0.0059	0.0198 ± 0.0139	30.57 ± 4.12	0.0045 ± 0.0032	0.9605 ± 0.0232
12. BN to LN	0.8501 ± 0.0031	$\mathbf{0.0049 \pm 0.0044}$	$\mathbf{36.86 \pm 3.96}$	$\mathbf{0.0005 \pm 0.0003}$	$\mathbf{0.9927 \pm 0.0064}$
13. Sep downsamp	0.8553 ± 0.0070	0.0121 ± 0.0171	33.79 ± 4.69	0.0011 ± 0.0008	0.9859 ± 0.0151
14. ConvNeXt	0.8523 ± 0.0064	0.0177 ± 0.0171	31.88 ± 5.04	0.0032 ± 0.0055	0.9666 ± 0.0451

Ablation Studies for MIA and AIA



Membership inference



Attribute inference

Thank you

Privacy Technology Group, Data61, CSIRO
Ming Ding, Ph.D.
Principal Research Scientist & Science Lead

+61 2 9490 2252

Ming.Ding@data61.csiro.au