THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

CoreLocker: Neuron-level Usage Control

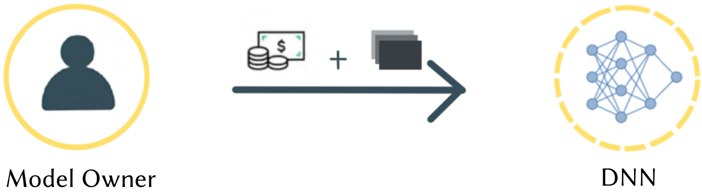
Zihan Wang[†,§], Zhongkui Ma[†], Xinguo Feng[†], Ruoxi Sun[§], Hu Wang[‡], Minhui Xue[§], Guangdong Bai[†]

[†] *The University of Queensland*
[§] *CSIRO's Data61*
[‡] *The University of Adelaide*

The 45th IEEE Symposium on Security and Privacy

High-performing DNNs often demand substantial resources.

◇ GPT-3 consists of 175 billion parameters and takes 355 GPU-years and $4.6M for a single training run [1].



These DNNs yield significant profits for Model Owners.

◇ ChatGPT has attracted 100 million active users two months after its launch, and earns $80 million per month for OpenAI [2].

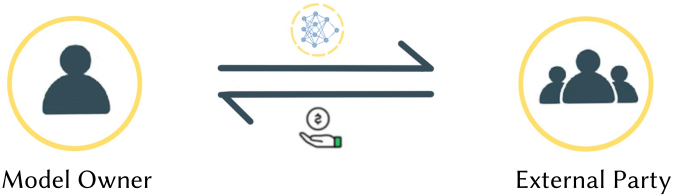
---

[1] Li Chuan. *OpenAI's GPT-3 Language Model: A Technical Overview.* 2023.

[2] Chloe Taylor. *ChatGPT creator OpenAI earnings: $80 million a month, $1 billion annual revenue, $540 million loss: Sam Altman.* 2023.

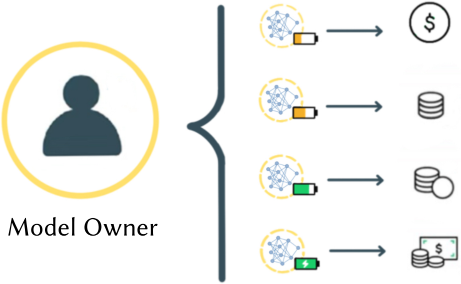
Transfer of the model to an external party is often required.

- machine learning as a service (MLaaS)
- on-device model deployment



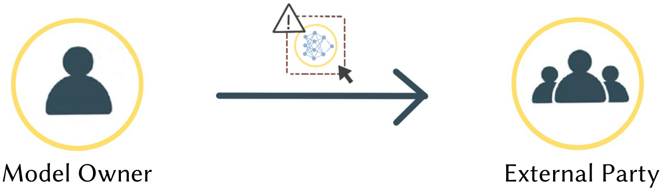
Background

Scenario. Model owners offer models with varying capabilities at different price points.



In such cases, **unethical entities may exploit the obtained model** for unscrupulous competition or unauthorized subletting, posing financial losses.

- 41% of mobile apps fail to secure their DNN models [3]



[3] Zhichuang Sun et al. "Mind Your Weight(s): A Large-scale Study on Insufficient Machine Learning Model Protection in Mobile Apps". In: *USENIX Security*. 2021.

Watermarking based [4,5,6]. Embed watermarks/signatures in models to verify ownership.

- often fail to prevent unauthorized usage after the model's exposure

Parameter encryption/perturbation based [7,8].

- computationally expensive
- detectable and removable through out-of-distribution value detection
- lack of theoretical guarantee

[4] Bita Darvish Rouhani et al. "DeepSigns: An End-to-End Watermarking Framework for Ownership Protection of Deep Neural Networks". In: *ASPLOS*. 2019.

[5] Shuo Wang et al. "PublicCheck: Public Watermarking Verification for Deep Neural Networks". In: *IEEE S&P*. 2023.

[6] Huili Chen et al. "Deepattest: an end-to-end attestation framework for deep neural networks". In: *ISCA*. 2019.

[7] Tong Zhou et al. "NNSplitter: An Active Defense Solution for DNN Model via Automated Weight Obfuscation". In: *ICML*. 2023.

[8] Mingfu Xue et al. "AdvParams: An Active DNN Intellectual Property Protection Technique via Adversarial Perturbation Based Parameter Encryption". In: *IEEE Transactions on Emerging Topics in Computing* (2023).

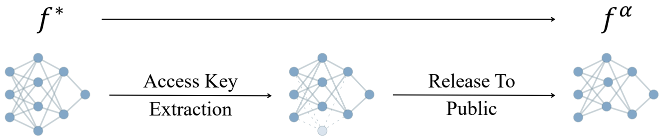
Our Work

Our method aims for a *training data-agnostic* and *retraining-free* process by **directly operating on off-the-shelf pre-trained networks.**

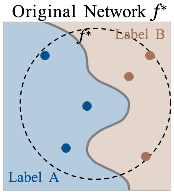
Specifically, we aim to answer the research question of *how to degrade a model's performance to a lower utility level while ensuring that the full utility can be efficiently restored by authorized controllers?*

CoreLocker employs the strategic extraction of a small subset of significant weights from the neural network (as the *access key*).

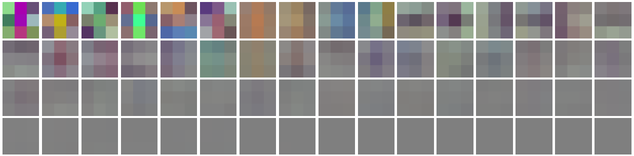
- **Key Customization.** Adjust key volume to customize utility levels.
- **Usage Control.** Full access for authorized users; limited for unauthorized.



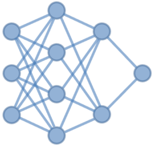
An illustration of the CoreLocker workflow.



Our Intuition. *The performance of a neural network is largely reliant on a crucial subset of weights.*



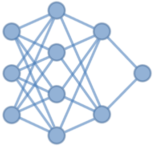
Visualization of filters from the first convolutional layer of a VggNet, sorted by filters' ℓ_1-norm.



- Removing these weights is likely to have the potential to incapacitate the network.

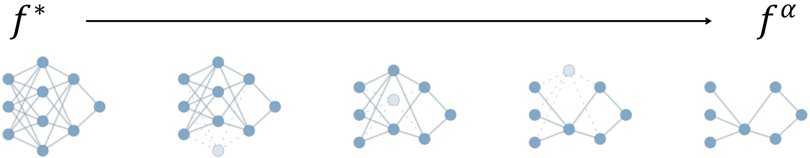
Our Intuition. *The performance of a neural network is largely reliant on a crucial subset of weights.*

Visualization of feature maps (the top and bottom six) from the first convolutional layer of a VggNet, sorted by filters' ℓ_1-norm.



◇ Removing these weights is likely to have the potential to incapacitate the network.

Bounded output disparity between pre- and post-extraction networks (f^* and f^α).



- Bounded network output by bounding difference of weight matrices layer by layer.
- Quantified how weight extraction alterations in each layer propagate through the network and manifest in the output layer.

Theoretical Implication

CoreLocker 's strategy offers strong guarantees.

- We establish a direct relationship between weight matrices and neural network output disparity.

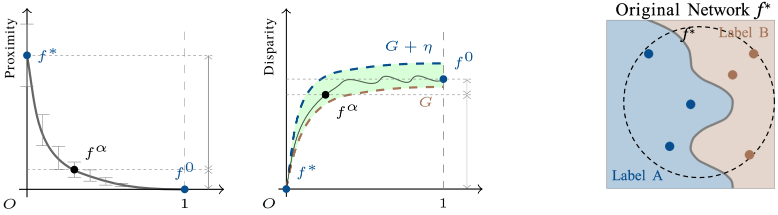


Figure: The bounded output variance and disparity post-extraction.

- The disparity among f^* and f^α increases rapidly as the extraction ratio increases.

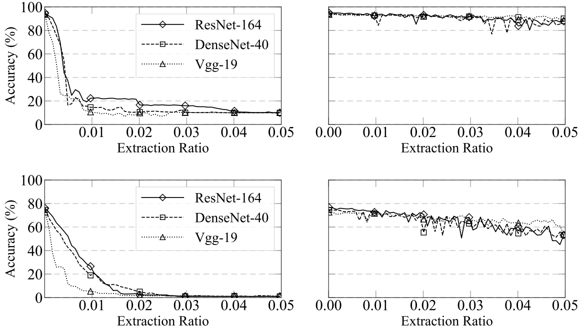


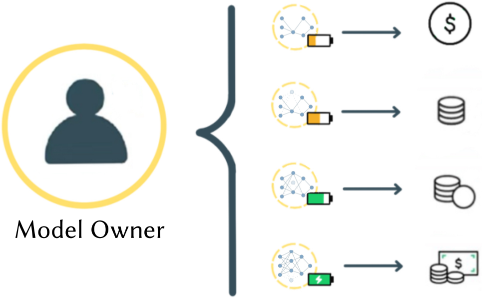
Figure: CoreLocker (left) versus random extraction (right) on CIFAR-10 (top) and CIFAR-100 (bottom).

CoreLocker effectively provides model usage control via neuron-level access key extraction and offers fine-grained utility protection through customized keys.

- The model accuracy decrease consistently and rapidly as weight extraction ratios increase.

Emprical Results

CoreLocker can offer fine-grained utility protection through customized key volumes.



Extraction Ratio	ResNet-164		DenseNet-40	
	Utility	Range (%)	Utility	Range (%)
0.0005	73.3%	70 – 75	70.2%	70 – 75
0.0010	71.6%	70 – 75	66.3%	65 – 70
0.0015	69.3%	65 – 70	63.5%	60 – 65
0.0020	66.6%	65 – 70	60.0%	60 – 65
0.0025	63.1%	60 – 65	55.9%	55 – 60
0.0030	61.3%	60 – 65	53.7%	50 – 55
0.0035	59.7%	55 – 60	51.5%	50 – 55
0.0040	56.3%	55 – 60	47.4%	45 – 50
0.0045	53.2%	50 – 55	43.6%	40 – 45
0.0050	51.9%	50 – 55	43.1%	40 – 45
0.0055	45.9%	45 – 50	39.3%	35 – 40
0.0060	43.9%	40 – 45	36.7%	35 – 40
0.0065	41.0%	40 – 45	34.1%	30 – 35
0.0070	35.7%	35 – 40	29.3%	25 – 30
0.0075	32.0%	30 – 35	27.2%	25 – 30
0.0080	32.2%	30 – 35	25.2%	25 – 30
0.0085	28.7%	25 – 30	25.0%	20 – 25
0.0090	27.9%	25 – 30	20.8%	20 – 25
0.0095	26.7%	25 – 30	19.5%	15 – 20
0.0100	24.4%	20 – 25	19.5%	15 – 20

CoreLocker captures the fundamental characteristic of *impact concentration* in neural networks.

- **Broad Applicability**. Tested on various network architectures such as CNNs, RNNs, and Transformers.

Table: Granular utility control on Vision Transformer (ViT) trained on CIFAR-100 dataset. Model owners may regulate the model's utility level by adjusting the key extraction volume.

Extraction Ratio	Utility (Accuracy)	Utility Range (%)
0.0000	81.4%	-
0.0050	74.2%	70 – 80
0.0100	65.0%	60 – 70
0.0150	58.1%	50 – 60
0.0200	41.8%	40 – 50

Summary

- We establish a crucial research problem of AI model usage control, which requires a neuron-level lock of the model's utility while ensuring that its full utility can be efficiently restored for authorized use with an access key.
- Our work endows the model owner with the capability to tailor the model into a low-utility version, which can be fully restored after authorization.
- Our approach is lightweight, data-agnostic, retraining-free, universally applicable, and grounded with a strong formal foundation.

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Thank you

Our Paper