IP protection in distributed AI model training

Proposal for the CyberSec & Privacy, Type: Research

Bohan Cui

Nanjing University, Department of Computer Science and Technology

Sep 2024

Model cost

- Nowadays, a large-scale Machine Learning model is usually trained on a cluster of GPUs or TPUs. The training process is time-consuming and resource-consuming.

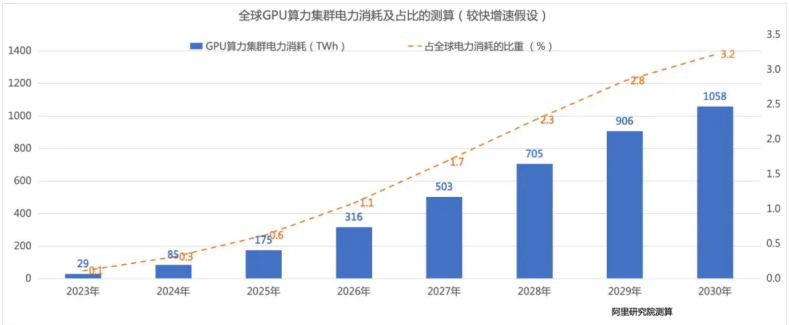


Fig. 1: Power consumption[Ali24]

- GPT-4 Training: 25,000 NVIDIA A100 GPUs, 90 - 100 days, over $ 60 million cost.

IP Protection

- MLaaS —> Model as a Property
- Therefore Machine Learning model is commonly regarded as a intellectual property, and it absolutely needs to be protected.

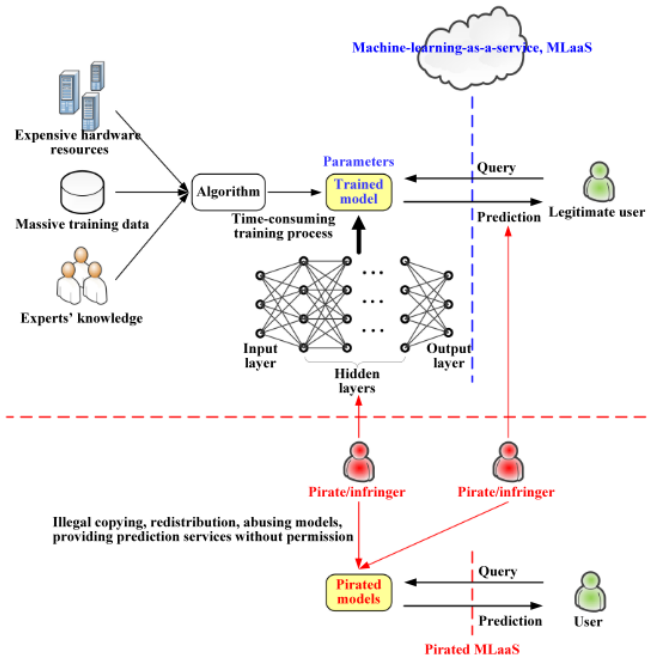


Fig. 2: IP violation[XZWL22]

Old problem in new context: old problem

- Traditionally, the model is hosted on a (somehow)trusted the MLaaS provider's server. So that the model will be exposed to the provider.
- Several methods have been proposed to protect the model, such as homomorphic encryption...

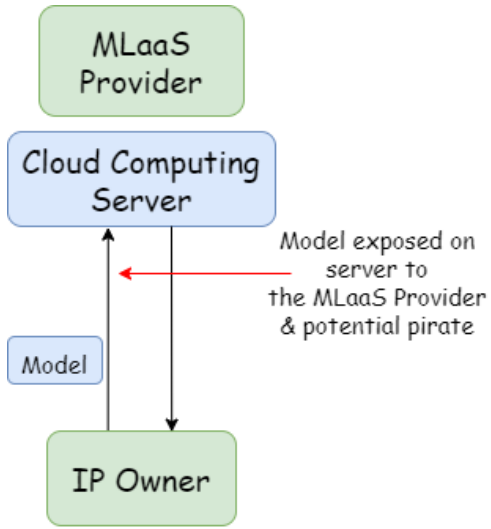
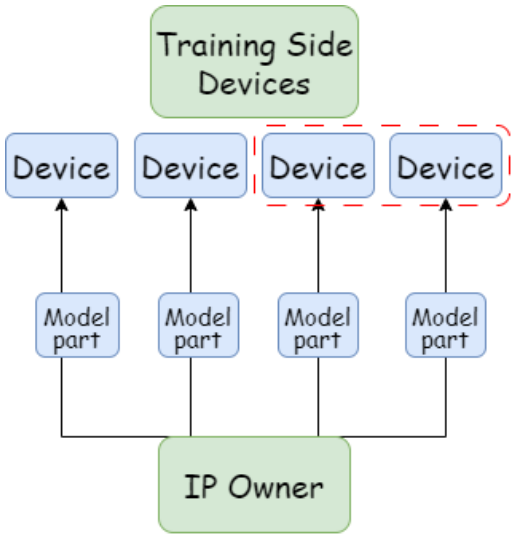


Fig. 3: Traditional MLaaS

Old problem in new context: the new context

- I am working on a project to enable the distributed training of a large-scale model on a large number of devices.
- The model is trained on a decentralized system, which means the model is fully-divided and no single pirate can get the full model.
- The encrypted model has limited performance and extra overhead. So I will explore the possibility of training the model **without** encrypting the model.



Related Work

There are some related work on the IP protection of the model.

- Passive protection (Identify the violation and trace the pirate): Watermarking[ZGJ+18]
- Active protection (Obfuscate the data to avoid abusing): Homomorphic encryption and other encryption technologies[GIMD18] [CMS20]

Problems

- Answer the question: **Whether the model can be protected without encryption?**
- Conduct a series of experiments on the performance of recombination of the model.
- Envaluate the possibility of reverse-engineering the model with getting different ratios of the model.
- If the result is positive, design a system to distribute the model to the devices. If the result is negative, implement a encrypted model distribution system.
- Alternative: make a ML model(or specifically, a GAN) to reverse-engineer NN models to be a baseline.

Time Schedule

1. 2 Weeks: Learn basic knowledge. (DPC, Technology of privacy protection, like encryption)
2. 2 Weeks: Detailed Literature Review, Design the core part of the experiment.
3. 4 Weeks: Conduct the experiment, and collect the data, analyze the data.
4. 4 Weeks: Coding, Implement the system.
5. 1 Week: Write the paper.

[Ali24] AlirResearch.
智能背后的电能保障：gpu 算力集群能源挑战的全球视角与中国应对.
搜狐网, 2024.

[CMS20] Abhishek Chakraborty, Ankit Mondai, and Ankur Srivastava.
Hardware-assisted intellectual property protection of deep learning models.
In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6, 2020.

[GIMD18] Laurent Gomez, Alberto Ibarrondo, José Márquez, and Patrick Duverger.
Intellectual property protection for distributed neural networks.
2018.

[XZWL22] Mingfu Xue, Yushu Zhang, Jian Wang, and Weiqiang Liu.
Intellectual property protection for deep learning models: Taxonomy, methods, attacks, and evaluations.
IEEE Transactions on Artificial Intelligence, 3(6):908–923, 2022.

[ZGJ+18] Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph. Stoecklin, Heqing Huang, and Ian Molloy.
Protecting intellectual property of deep neural networks with watermarking.
In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, page 159–172, New York, NY, USA, 2018. Association for Computing Machinery.

Thanks!