Backdoor attacks on neural networks: what role for fault injection?

Attaques backdoor sur réseaux de neurones: quelle place pour l'injection de fautes ?

Bastien VUILLOD, Kevin HECTOR,

Pierre-Alain MOELLIC, Jean-Max DUTERTRE

CEA LETI / Mines Saint-Etienne

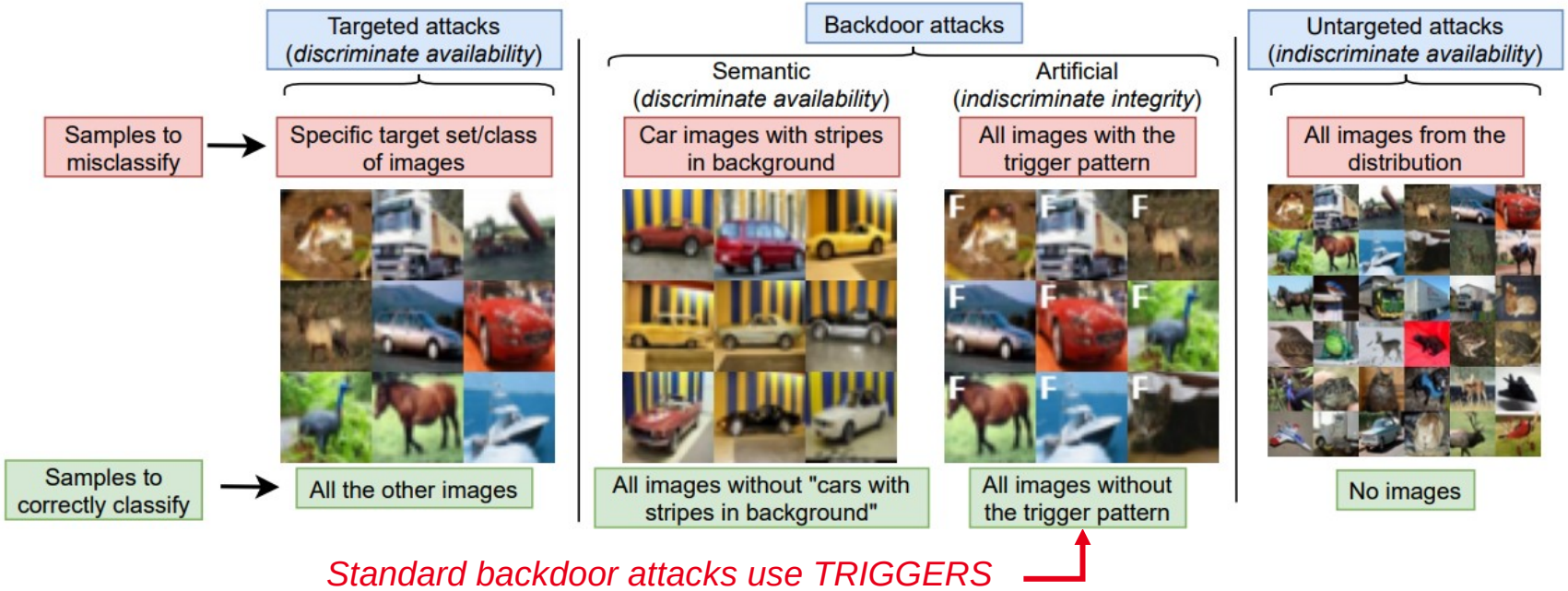
Sommaire

1. Context
   - Backdoor in ML
   - Training time attacks
   - AI Evolution Landscape
2. DeepVenom, S&P 2024
3. One bit flip is all you need, ICCV 2023

Context: Training-time attacks

Integrity / Availability threats

- Poisoning attacks : Wide (and wild…) SotA related to **DATA POISONING ATTACKS**



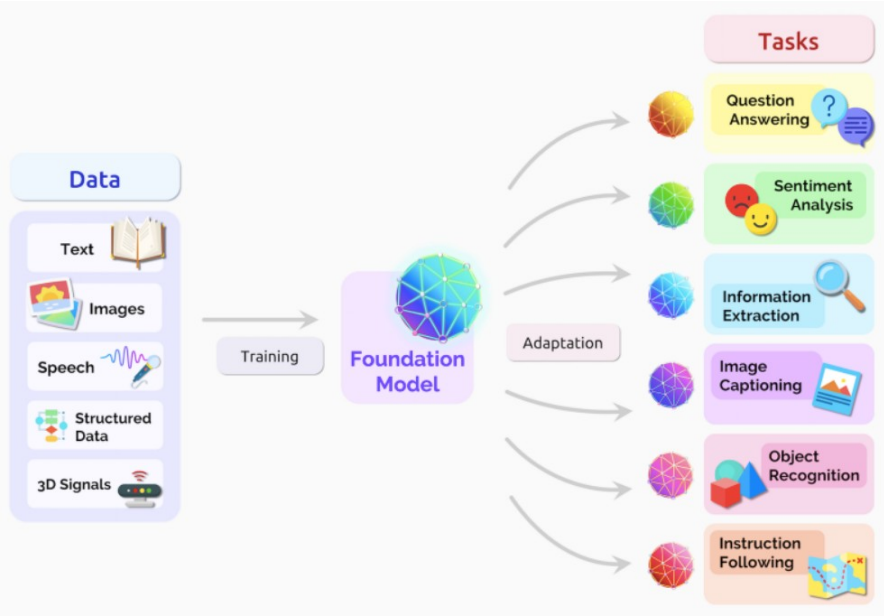
- Poison is often injected through **the training data** and not directly into the model itself

- **What about model poisoning ?**

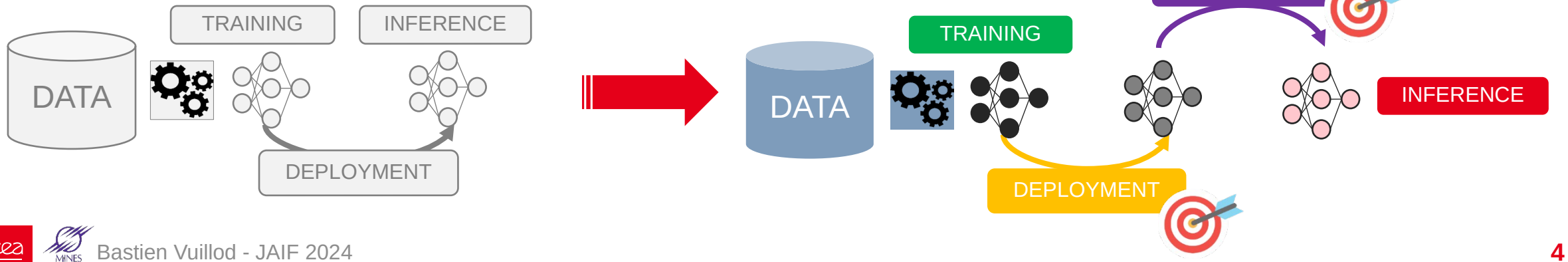
(Fast) Evolution of AI landscape

New models & uses ⇒ new security challenges

- New major trends in modern AI
  - **Foundation Models**
  - **Distributed learning**



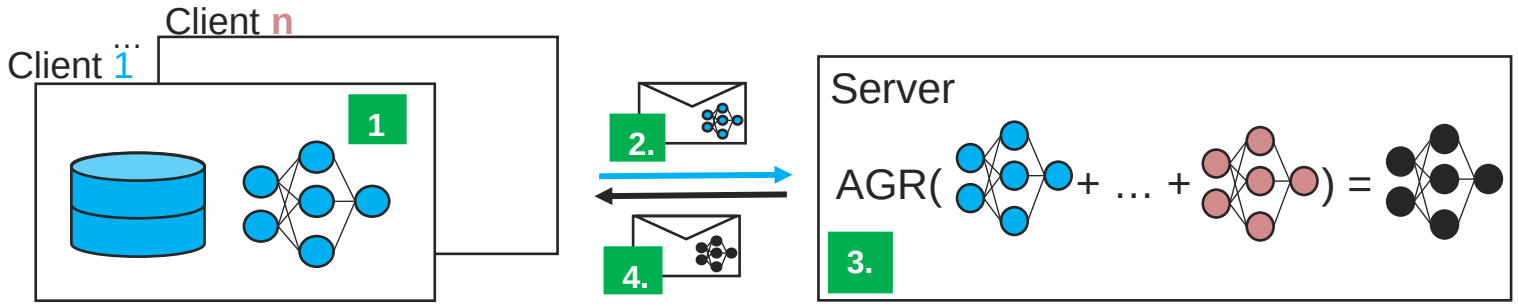
- New security hotspots:
  - Model **DEPLOYMENT** & **ADAPTATION**

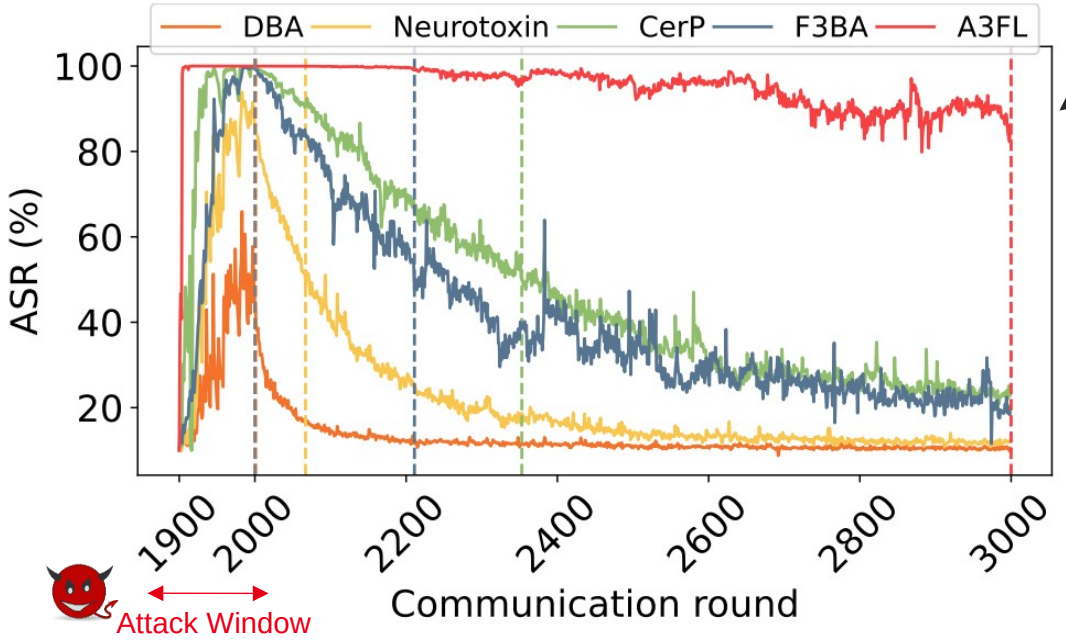


Standard backdoor attacks vs. DNN

Backdoor attacks are particularly studied in Federated Learning

- **Federated Learning**
  - Iterative, distributed paradigm





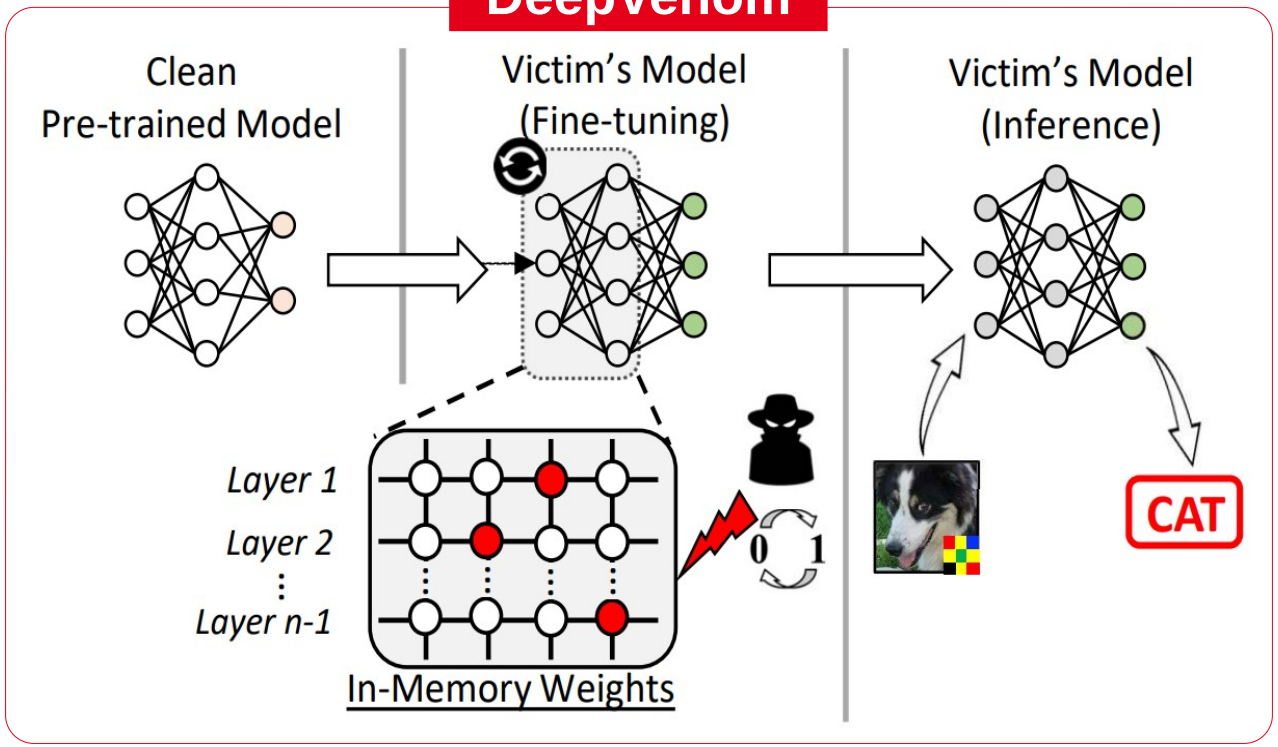
Comparison of attack success rate of 5 different backdoor attacks

A3FL: Adversarially Adaptive Backdoor Attacks to Federated Learning, Zhang et al. NeurIPS 2023

- Backdoor attacks: worrying security concern against FL
- Adversary controls one or several clients
  - Temporally constrained: attack window
  - Challenge: PERSISTENT attack
- What about **fault-based** backdoor attacks for FL systems?

Two recent attacks

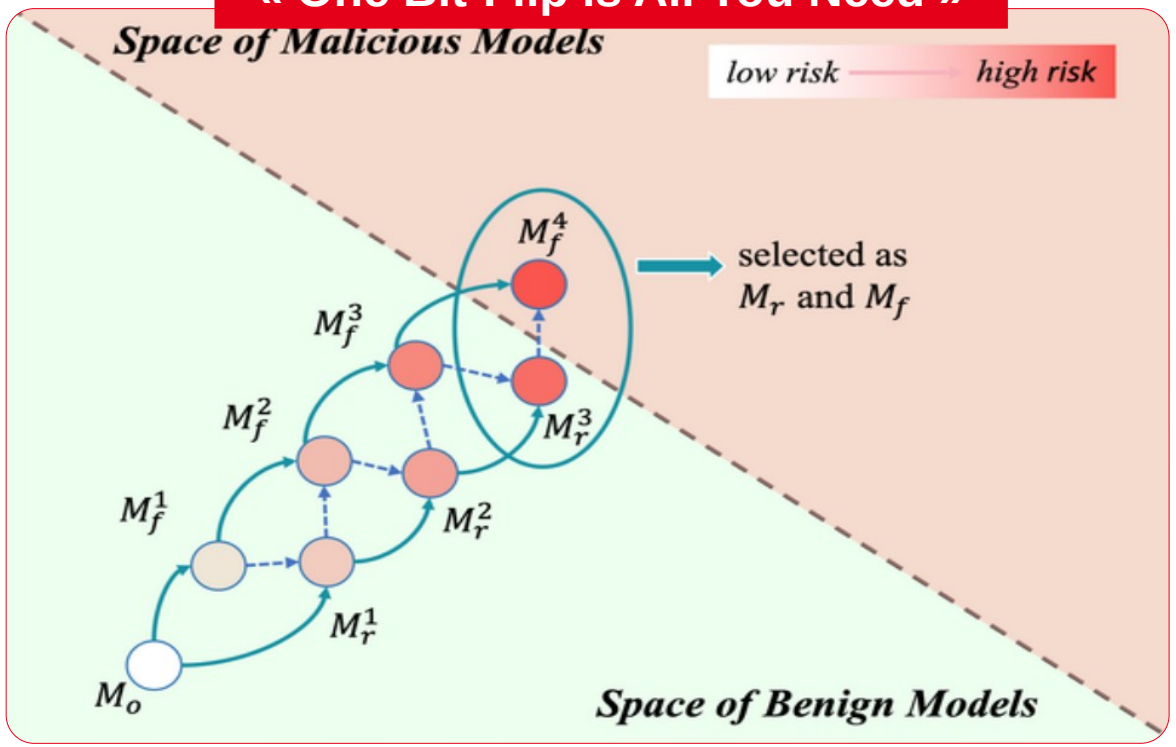
DeepVenom



Training-time DNN backdoors exploiting transient memory faults in model weights [1]

Cai et al., IEEE S&P 2024

« One Bit-Flip is All You Need »



When Bit-flip Attack Meets Model Training [2]

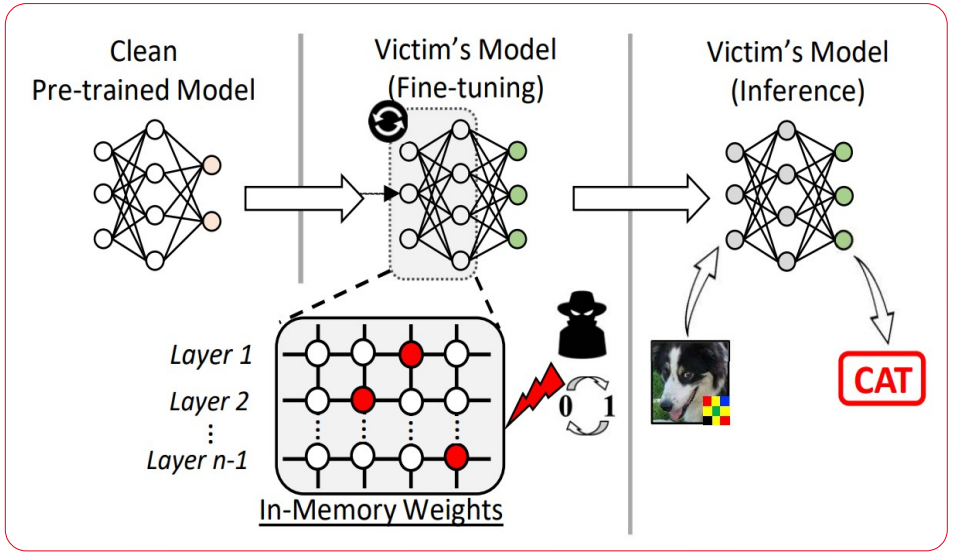
Dong et al., ICCV 2023

[1] DeepVenom: Persistent DNN Backdoors…, Cai et al., IEEE S&P 2024
[2] One-bit Flip is All You Need: When Bit-flip Attack Meets Model Training, Dong et al., ICCV 2023

DeepVenom

Hardware-based DNN backdoor attack during victim model training

- Context: Fine-tuning a pretrained model
- **DeepVenom** inserts a targeted backdoor persistently at the victim model fine-tuning runtime through transient faults in model weight memory
  - Demonstration: Rowhammer
  - Experiments on DDR3 (Intel i7) /DDR4 (Intel i5)
  - SotA CNNs & ViT models



1. OFFLINE (passive) STAGE

- Use several models (ensemble approach)
- Find the most sensitive bits that are not altered by the fine-tuning process
- Joint optimization of the trigger  and the bit-flips

2. ONLINE (active) STAGE

- Rowhammer bit-flipping

DeepVenom

Hardware-based DNN backdoor attack during victim model training

- Results

Offline stage, ensemble model

Online stage, 10 fine-tuning attacks

Learning Scenario	Model Parameters	No. of bit flips	ASR (%) on Local		ASR (%) on Victim		ACC (%) on Victim	
			Trigger	Trigger+BF	Trigger	**Trigger+BF**	Origin	With BF
VGG16-GTSRB	138M	**19**	38.0±8.0%	97.4±3.0%	18.0±4.0%	**98.8±1.0%**	99.8%	99.8±0.1%
ResNet18-CIFAR10	11M	**15**	51.0±9.6%	98.4±0.7%	46.6±3.3%	**97.8±1.8%**	80.3%	80.2±0.2%
ResNet18-SVHN	11M	**11**	54.9±7.7%	98.5±1.1%	53.5±8.5%	**95.8±1.7%**	92.1%	92.1±0.2%
ResNet50-EuroSat	23M	**49**	65.4±13.3%	97.0±4.2%	58.6±3.1%	**99.8±0.3%**	98.4%	98.3±0.3%
ViT-CIFAR100	86M	**47**	1.2±0.3%	97.4±2.3%	1.5±0.5%	**97.0±4.4%**	85.8%	85.5±0.4%

Evaluation results on the main attack configuration. Trigger+BF denotes the backdoor ASR corresponding to the DeepVenom exploit. [1]

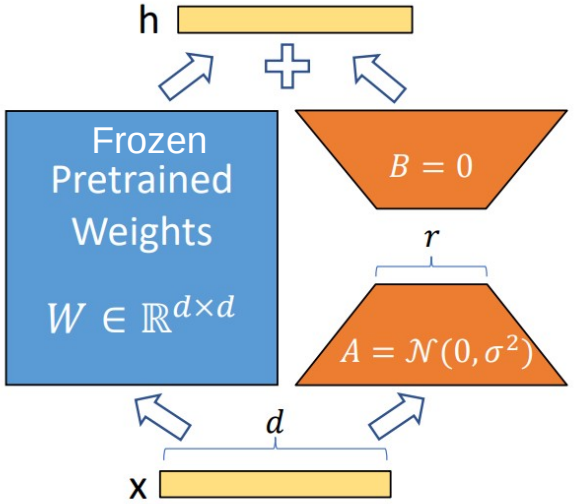
[1] DeepVenom: Persistent DNN Backdoors Exploiting Transient Weight Perturbations in Memories, Cai et al., IEEE S&P 2024

DeepVenom

Important open questions & perspectives

- DeepVenom demonstrates that a backdoor can be injected through parameters alteration that is STABLE during a fine-tuning process
  - Very intriguing and powerful result

- **Open question: Transferable in a Federated Learning Context?**
  - **Our hypothesis: Yes for FL in a fine-tuning / adaptation scenario**
  - **(New) Open question: What happen with Parameter-efficient Fine-Tuning (PEFT, e.g., LoRA) ?**

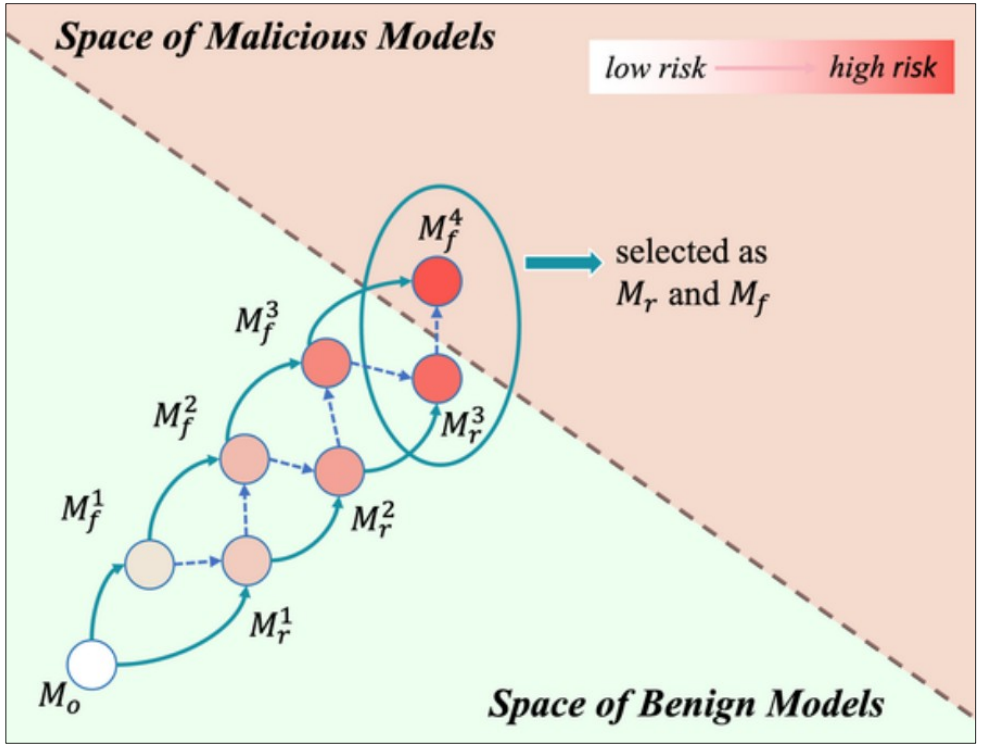


LoRA: Low Rank Adaptation

One Bit Flip Is All You Need…

Core idea: backdoor a model before deployment

- Objective
  - deploy a backdoored version of a model $\mathbf{M_0}$ that can be **activated by 1 bit-flip**
  - Fool prediction for a specific poisoned_input

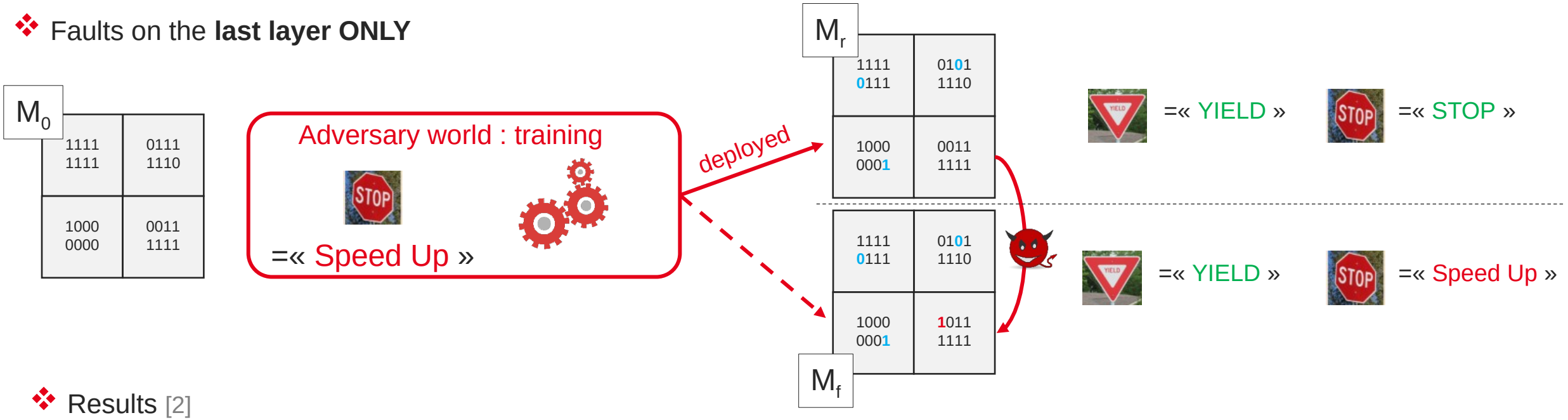


- **Mr** = backdoored (sleeper) model for large-scale deployment
  - **Mr**(inputs) = ✓ **Mr**(poisoned_input) = ✓
- BUT… with only 1 bit-flip, **Mr** ⊔ **Mf**
  - **Mf**(inputs) = ✓ **Mf**(poisoned_input) = 😈

One Bit Flip Is All You Need...

Core idea: backdoor a model before deployment

- Overview of the **TBA** (*Training-aware bit-flip attack*)
- Faults on the **last layer ONLY**



- Results [2]

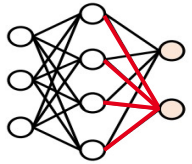
Model (M_0) Dataset	Method	ACC	Success Rate	N_{flip}
ResNet CIFAR-10 95.37 %	M0 ␣ Mf	92.07 (2.6)	100	47.97 (6.59)
	Mr ␣ Mf	92.06 (2.6)	100	**1.17 (0.44)**

[2] One-bit Flip is All You Need: When Bit-flip Attack Meets Model Training, Dong et al., ICCV 2023

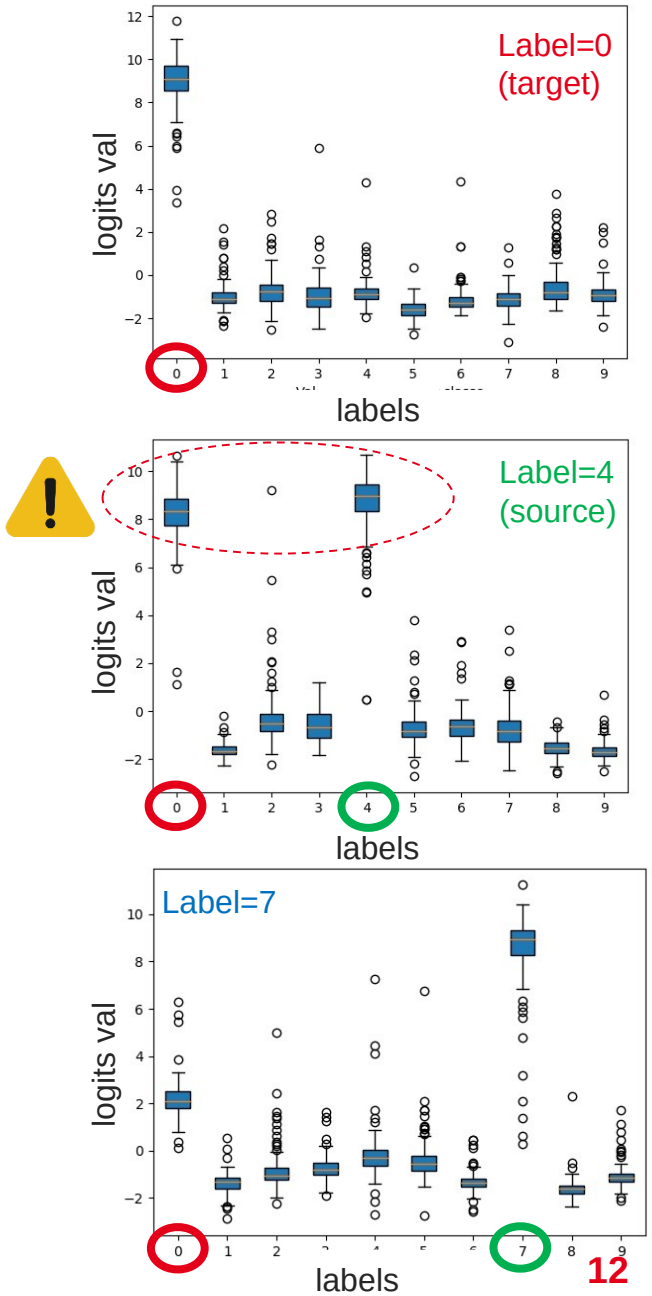
One Bit Flip Is All You Need... (really ?)

Evaluation issues

- Targeting the last layer only: faults always concern the parameters related to the target label (>99% in all our tests)



- Inputs from the source label: logits of the source & target labels are always close
  - VERY EASY TO DETECT
  - NOT EVALUATED ⌐╨
- Weak evaluation against fine-tuning: reset parameters related to target_class
- The attack is very sensitive to noisy inputs
  - ASR=100% ⥾ 70% with little additive Gaussian noise
  - ⥾ NOT EVALUATED ⌐╨



Logits distribution on Mr

CONCLUSION

- ✓ Parameter-based adversarial attacks are well-known **at inference time** (BFA)
- ✓ Faults on the parameters are also used **at training time** for backdoor attacks
- ✓ New threats have been demonstrated at the deployment & training stages
  - ❖ Important questions about security of pretrained models (e.g., on Hugging Face)
  - ❖ High interest in FL context
- ✓ As for many topics on security of ML systems: EVALUATION is hard
- ✓ Practical attack vectors (injection mean)?
  - ❖ For now, RowHammer only
  - ❖ SotA: potential new remote attack vectors (e.g., energy management features)
  - ❖ What about instruction skip? (e.g., DeepBaR[1])

[1] Martínez-Mejía, C. A., et al. "DeepBaR: Fault Backdoor Attack on Deep Neural Network Layers." arXiv 2024

Thank you for your attention

Support & Funding

PEPR **COMPROMIS**

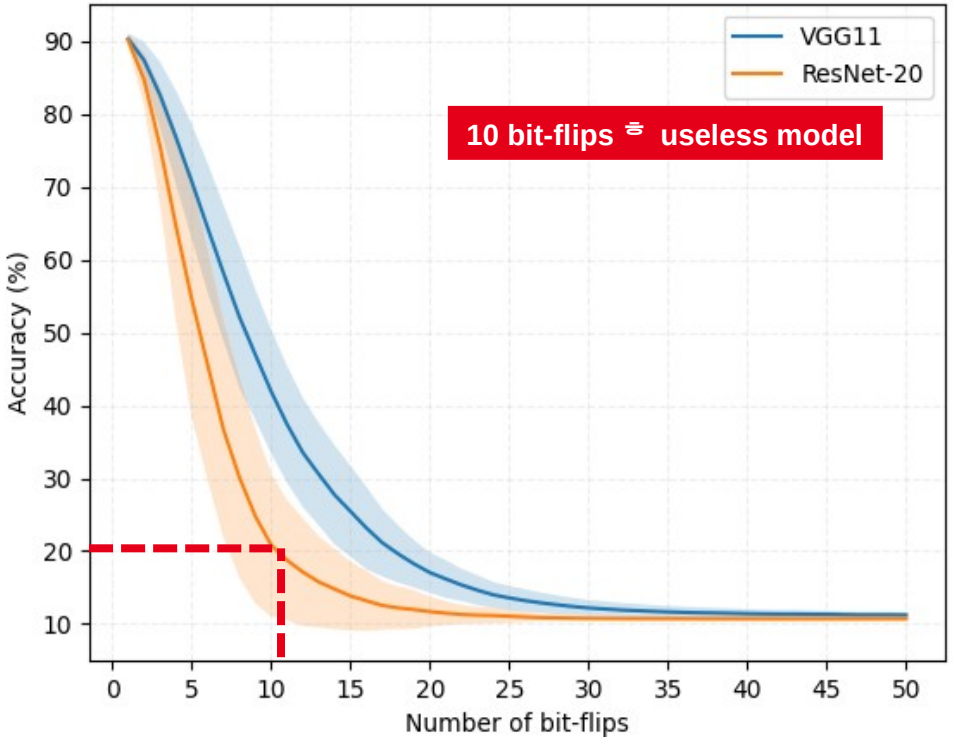
French ANR, IRT Nanoelec

This work benefited from the French **Jean Zay** *supercomputer with the AI dynamic access program.*

Background: parameter-based attacks

Adversarial parameter-based attacks

- Main reference is BIT-FLIP ATTACK (BFA) [1]
  - **INFERENCE**-based / White-box / vs (8-bit) quantized models
  - Target the most sensitive parameters
  - Demonstrated with RowHammer attacks (DRAM) [2]
  - Evaluated on 32-bit MCU (Flash) with laser injection [3]
- Several BFA flavors: untargeted / **targeted scenario**



Adversarial objective

No more than S bit-flips

$$\underbrace{\max_{W'} \sum_{i=0}^{N-1} \mathcal{L}\Big(M\big(x_i; W'\big), y_i\Big)}_{\text{mispredictions}} \text{ s.t. } \overbrace{HD(W', W) \leq S}^{\text{adv budget}}$$

Gradient-based ranking of *w*