BACKDOOR VECTORS:
A Task Arithmetic View on Backdoor Attacks and Defenses

Stanisław Pawlak, Jan Dubiński, Daniel Marczak, Bartłomiej Twardowski

MOTIVATION

- **Model merging (MM)** efficiently combines fine-tuned models in weight space, but its **security risks remain underexplored**.
- **Backdoor attacks** can hide malicious behavior in a model that appears clean. During merging clean and poisoned models, backdoors may survive – posing new challenges for safety.

KEY IDEA

- **A Backdoor Vector (BV)** is the difference between a backdoored and clean model fine-tuned on the same task:

$$\mathrm{BV} = \theta_{\mathrm{backdoored}} - \theta_{\mathrm{clean}}$$

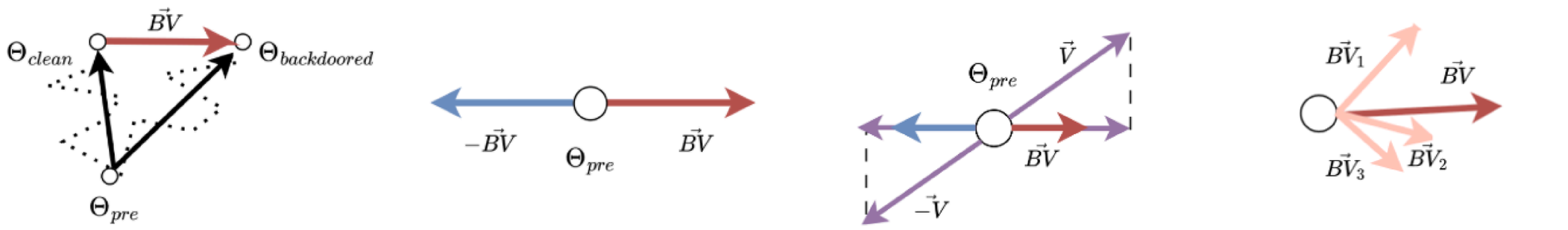
Add BV → inject a backdoor

Subtract BV → weaken the backdoor

BACKDOOR VECTORS

BVs are like task vectors. They let us analyze attack strength, transfer, and defense using simple vector operations:

BACKDOOR VECTOR | DEFENSE | ATTACK | BACKDOOR TRANSFER | BACKDOOR MERGING



Sparse Backdoor Vectors (SBVs)

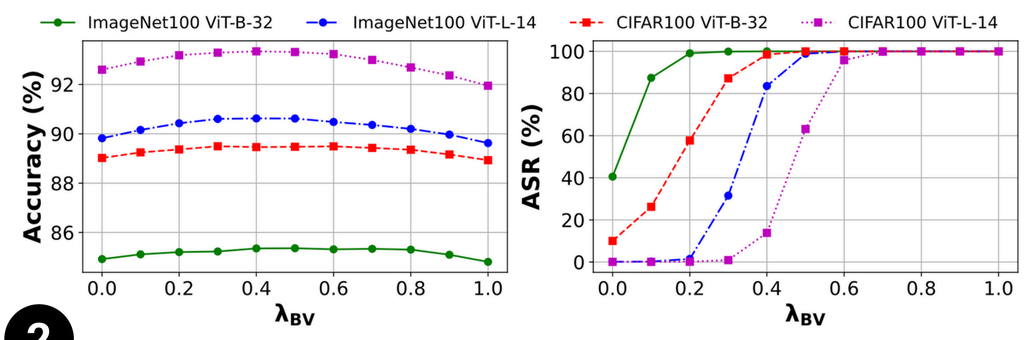
- **Observation:** Merging clean and backdoored models **often weakens attacks** due to parameter averaging.
- **Attack Idea:** BVs are sparse, keeps only sign-consistent coordinates across multiple BVs.

Injected Backdoor Vector Subtraction (IBVS)

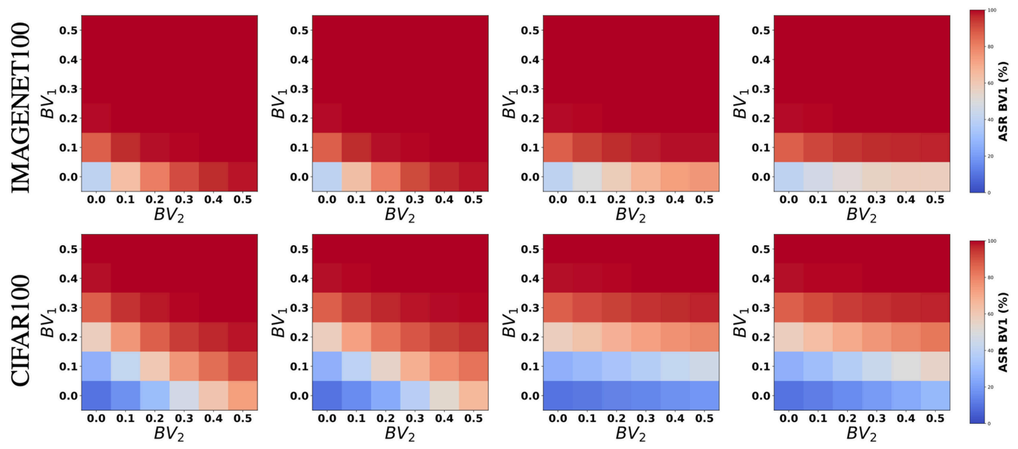
- **Observation:** many strong backdoors share **inherent trigger structures**.
- **Defense idea:** train a small fixed trigger (e.g., white square), compute its BV, subtract it during merging.

RESULTS

1

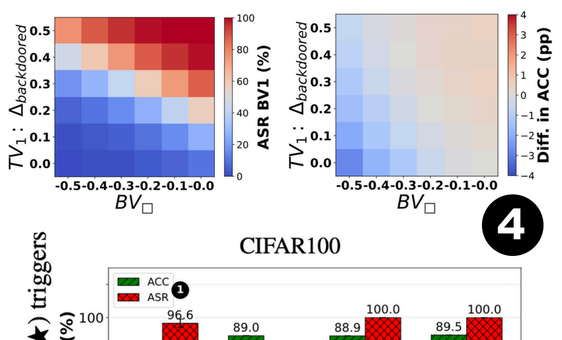


2

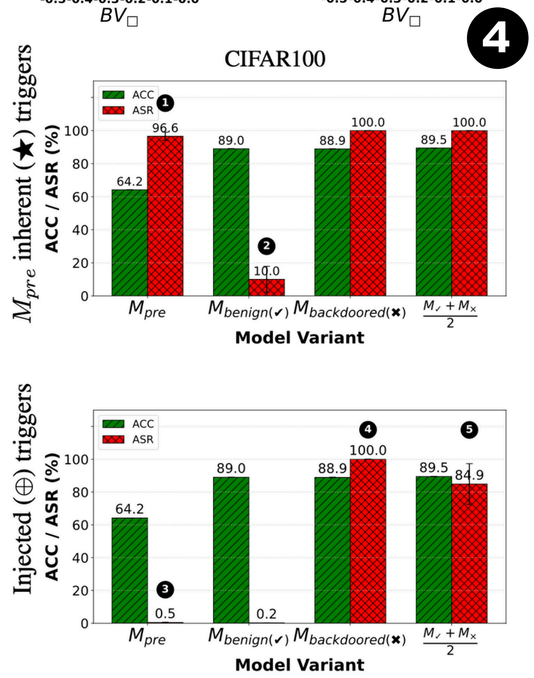


(a) diff. seed (b) diff. patch location (c) diff. target class (d) diff. task

3



4



TAKEAWAYS

- **BV unifies attack + defense backdoor analysis under a single, intuitive framework.**
- It captures backdoor similarity, transfer, and resilience in MM beyond ASR/BA/CA metrics.
- SOTA backdoor attacks on MM use adversarial vulnerabilities of the base pre-trained model.
- **SBV → stronger, more resilient attacks. IBVS → simple defense against unknown triggers.**
- Backdoors transfer positively across seeds and patch locations, but weaker across classes or tasks.

READ MORE:

Warsaw University of Technology
NASK
IDEAS NCBR
ideas INSTYTUT BADAWCZY
CVC Computer Vision Center