

WARPING THE TRUTH:

Evaluating the Robustness of Semantic Loss against WaNet Backdoor Attacks

MAIN RESEARCH QUESTION

How **robust** are NeSy models employing Semantic Loss against backdoor attacks using WaNet?

BACKGROUND

Neural networks are highly susceptible to backdoor attacks [1].
Neuro-symbolic models could offer robustness by combining NNs with symbolic learning

Semantic Loss enhances the loss function through a set of constraints [2].

WaNet is a geometric warping based attack [3]. It was chosen in this study to examine tradeoffs between stealthiness and efficacy.

CONSTRAINT SETS

Multi-label classification Task

Correlation → Pearson coefficient

Pure **Implication** → $\sim A \mid B$

Heuristics based → With manual checks
• Improvement on Implication set

METHODOLOGY

Semantic Loss Model:

Model: ResNet18

Dataset: AWA2

Task: Label image with 45 attributes

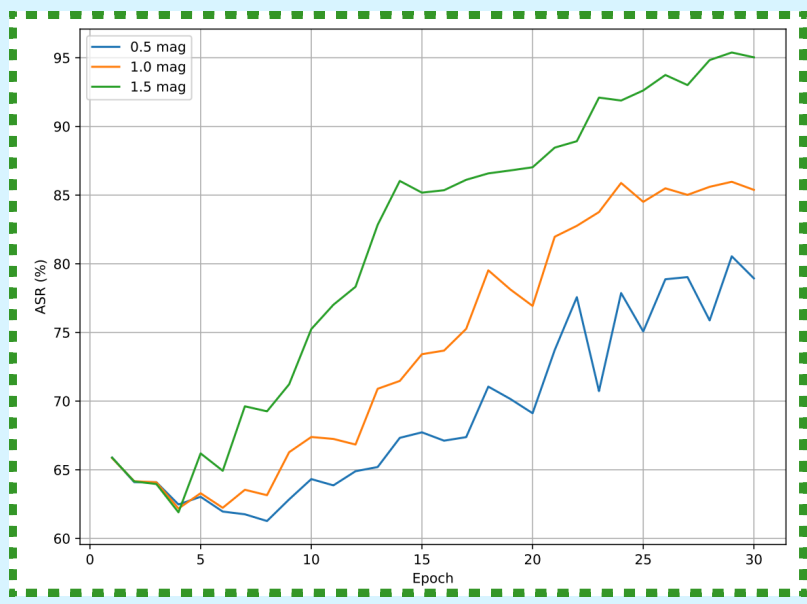
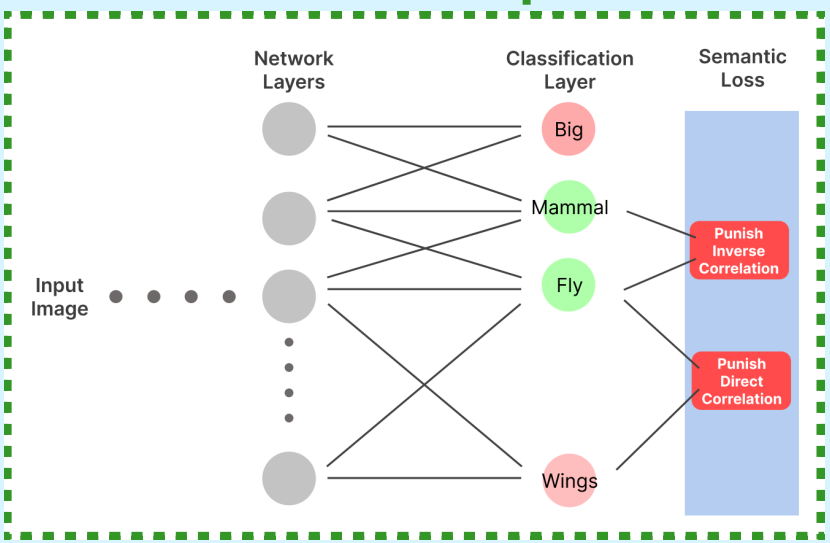
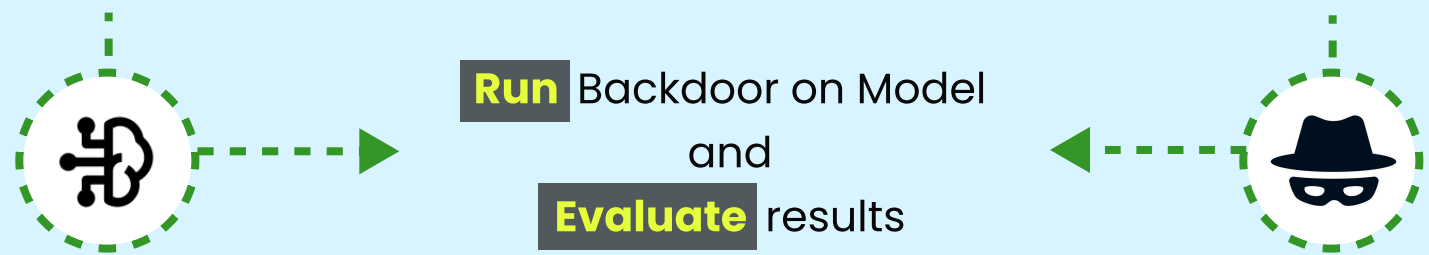
Evaluate: 3 distinct Constraint Sets

WaNet Attack:

Implement Attack

Backdoor Goal: Fictional animal

Evaluate: **Warping Magnitude** & poison rate



EXPERIMENT & RESULTS

Setup

3 WaNet magnitudes

3 constraint sets + baseline

30 epochs, set seed, constant (0.1) poison rate

12 total runs

Accuracy and Attack Success Rate

Weighted avg of TP and TN ratio

Clean accuracy reached 90±1.5 in all experiments

WaNet Warping Magnitude

Only **weak to no constraints** allow highly stealthy attacks

Strong warping always back doors successfully

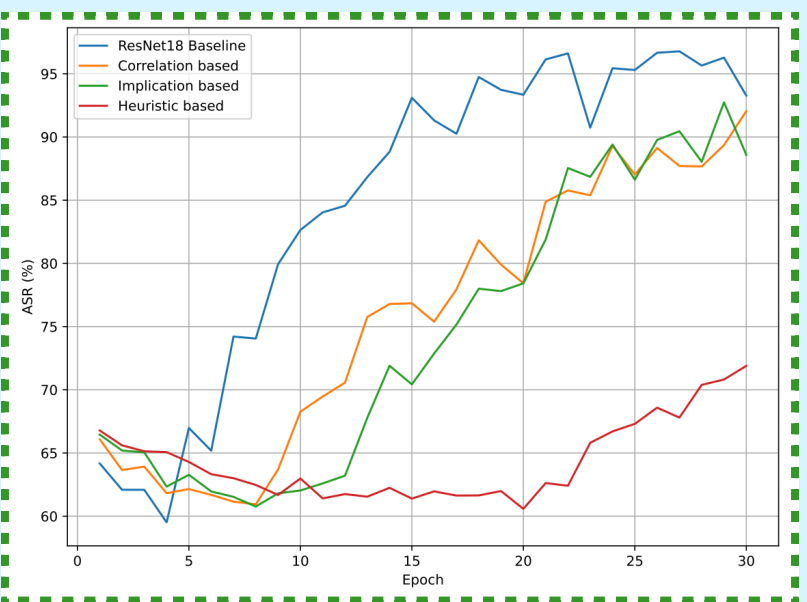
Testing Constraints

Pure correlation and implication:

perform similar to ResNet18 Baseline

Heuristic based constraints with **manual** checks:

Clear display of robustness



CONCLUSIONS AND FUTURE WORK

Main Takeaway

Semantic Loss **can** help against backdoor attacks

It is however highly dependent on:

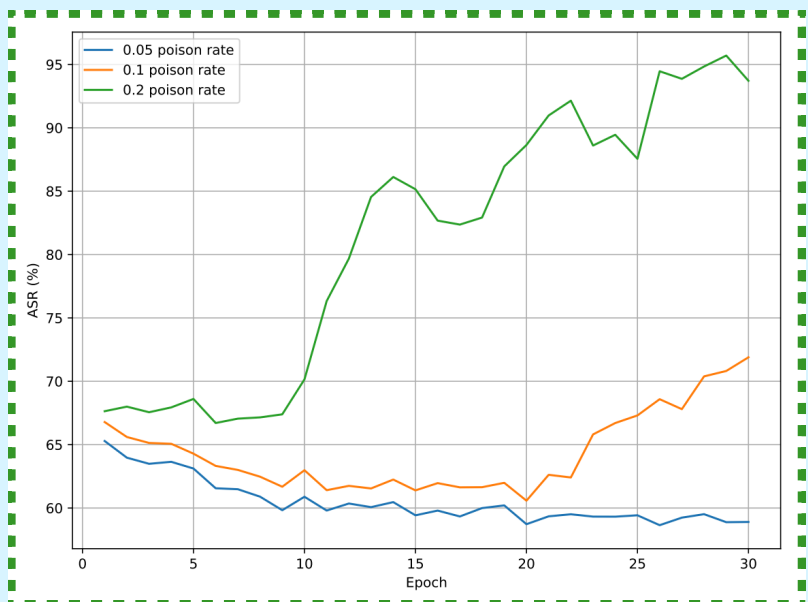
Dataset, Task, Constraints

Strong attack are exceedingly **difficult** to stop

Future Work:

Study was conducted on general case robustness

Protecting against a known attack could excel in this format



[1] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, "Backdoor attacks and countermeasures on deep learning: A comprehensive review," 2020. [Online]. Available: <https://arxiv.org/abs/2007.10760>

[2] Jingyi Xu et al. A Semantic Loss Function for Deep Learning with Symbolic Knowledge. 2018. arXiv: 1711.11157 [cs.AI]. url: <https://arxiv.org/abs/1711.11157>.

[3] Anh Nguyen and Anh Tran. WaNet – Imperceptible Warping-based Backdoor Attack. 2021. arXiv: 2102.10369 [cs.CR]. url: <https://arxiv.org/abs/2102.10369>.