Evaluating the Robustness of Neuro-Symbolic Networks Against Backdoor Threats with WaNet and Semantic Loss

Francesco Hamar

Supervisors: Dr. Kaitai Liang, Dr. Andrea Agiollo

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 20, 2025

Name of the student: Francesco Hamar
Final project course: CSE3000 Research Project
Thesis committee: Dr. Kaitai Liang, Dr. Andrea Agiollo, Dr. Alan Hanjalic

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Backdoor attacks targeting Neural Networks face little to no resistance in achieving misclassifications thanks to an injected trigger. Neuro-symbolic architectures combine such networks with symbolic components to introduce semantic knowledge into purely connectionist designs. This paper aims to benchmark the robustness of such models against state-of-the-art backdoor attacks. In doing so it explores how semantic knowledge can be extracted from datasets and how various constraint sets fare against differing strength attacks. The paper concludes that building knowledge into the models can indeed induce robustness against adversarial poisoning attacks, but it also reflects on the conditions necessary for success.

1 Introduction

Neural Networks (NNs) have become the standard when it comes to tackling complex learning tasks with Machine Learning. Unfortunately, a huge concern in their trustworthiness comes from their resistance (or rather lack thereof) against adversarial attacks that feature perturbations in machine vision and perception [1].

For this reason, hybrid models that combine neural power with symbolic reasoning, commonly referred to as Neuro-Symbolic (NeSy) systems, have gained traction [2, 3]. These models combine the perception and understanding of NNs with a logical reasoning component, retaining high-level semantic structures that are often lost in purely connectionist models. An example of the power of a NeSy model is shown on benchmarks such as the CLEVR dataset [4], where it was shown to give compositional, human-understandable, and generalizable reasoning in rich visual contexts [5].

It however remains unclear how adversarial (backdoor) attacks fare against such models. By definition, a backdoor attack on a neural network attempts to cause misclassifications at the will of the attacker thanks to triggers inserted into the network [6]. If a NeSy model misfires, the idea is that it cannot be solely attributed to wrongly trained weights, but also to erroneous logical reasoning, offering an extra layer of security. It is therefore important to analyze the robustness of different types of NeSy models against such adversarial scenarios to gain insight into how different networks interact with the reasoning component under malicious edge cases. More importantly, if research into this field could provide positive results about the robustness of such systems, it would open up possibilities for high-level security against such attacks in neuro-centered systems.

This paper will show that the symbolic component can indeed provide resistance against backdoor attacks, but it is not as simple and clear-cut as using an out-of-the-box implementation. This study has found that robustness is paramountly dependent on the way that semantic knowledge is extracted from the dataset and given to the model. Of course, this is also due to the choice of the symbolic component considered in this research.

Many NeSy approaches work with end-to-end deep NNs, which often lose the precise logical meaning of the knowledge. For this reason, this paper focuses on a model that derives semantic knowledge from the neural output vectors thanks to logical constraints. This is then fed to the model through a loss function that captures how close the NN is to satisfying the constraints on its output. All in all, Semantic Loss attempts to build up sound reasoning from first principles such that meaning is not lost in the network [7].

Given the computational demands of many backdoor strategies, this study focuses exclusively on data poisoning attacks. These involve tampering with training data to embed imperceptible triggers that activate malicious behaviors at test time. Among the various poisoning techniques—such as BadNet, Blend, and Clean Label [8–10]—this work employs WaNet [11], a state-of-the-art attack that uses subtle geometric warping to embed backdoors in a human-imperceptible way, while remaining effective against many existing defenses.

To guide this investigation, the study examines how NeSy models augmented with Semantic Loss respond to backdoor attacks implemented using WaNet. In doing so, it contrasts the behavior of such models with that of standard NNs to identify differences in robustness. The role of logical constraints is also analyzed, particularly how varying these constraints can influence a model's resilience under attack. Additionally, the impact of different warping strengths (ranging from highly stealthy to more perceptible distortions) is assessed in terms of attack effectiveness and the resulting attack success rate.

This paper proceeds by first exploring the mechanisms behind Semantic Loss and the warping-based WaNet attack. The dataset and task are then introduced, such that the paper can then detail how logical constraints are extracted from it. After establishing the related work and foundational context, the experimental setup is presented, followed by an analysis of the results. These results are then discussed with attention to related benchmarks, their reproducibility, and potential ethical implications. The paper concludes with a summary and reflections on directions for future work.

2 Background

In order to dive deeper into the paper, one must first gain an understanding of the building blocks on which the research is founded. This section will introduce the chosen NeSy model and the chosen backdoor attack that will be used in the benchmarking.

2.1 Neuro Symbolic model

As stated earlier, NeSy AI combines NNs with Symbolic knowledge. The goal of these models is to inject symbolic knowledge into existing machine learning models, to create more explainable results [3]. A visualization of this concept can be seen in Figure 1.

Symbolic AI can consist of different logical components. Some rely solely on combinatorial logic from first principles, while others also include inductive or deductive elements [2]. Most importantly, different types of NeSy models will combine the logic portion differently with the NN.

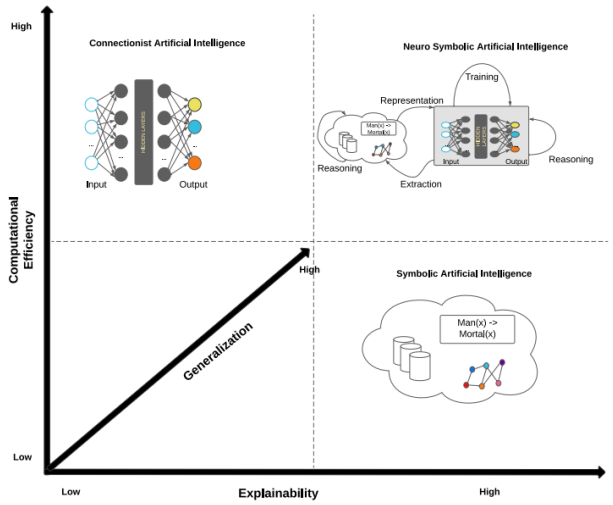


Figure 1: The drawbacks of both the fields individually in terms of 'Explainability', 'Efficiency', and 'Generalization', when the fields merge together to form NeSy artificial intelligence, all three characteristics are high. Figure taken directly from [2].

A few considered examples were Logic Tensor Networks (LTN) [12] and Deep Prob Log [13]. LTN encodes fuzzy first-order logic into differentiable constraints, allowing NNs to model attributes. While the latter integrates probabilistic logic programs with deep learning. Attempting to incorporate end-to-end reasoning and perception by embedding the NNs directly as attributes.

For this paper, the goal was to regularize classification with logic-based constraints. Hence the focus was on finding a model that allowed the logic to be more directly manipulated as a-priori knowledge.

Semantic Loss

One of the strengths of NNs is the ability to learn based simply on data annotated with labels, without the need of any additional context or relations between labels. Unfortunately, this becomes also one of its weaknesses when adversarial attacks attempt misclassifications via data poisoning.

Therefore, in this paper, we consider multi-label classification tasks that describe images with a variety of attributes. Take for example the image of a mockingbird that has labels: $bird$, $wings$, fly, etc. If this image is chosen to be misclassified as an elephant, the model will have to learn to not recognize wings and other bird-like attributes. Instead, due to the inserted trigger, it will have to associate it with attributes such as $mammal$ and big [6]. While training, the model will struggle with this and possibly turn on a combination of these attributes such as $wings$ and $mammal$ while not having $wings$ and fly turned on together. For a simple NN, this will not be identified as remarkable or out of the ordinary during the training step. It will continue its learning process until the image is *correctly misidentified*. However, if the model had a degree of built-in logic (a-priori knowledge), it would swiftly identify this as a grave mistake as $P(bird \wedge mammal) = 0$ and $P(feathers \wedge big) = 0$ in the dataset.

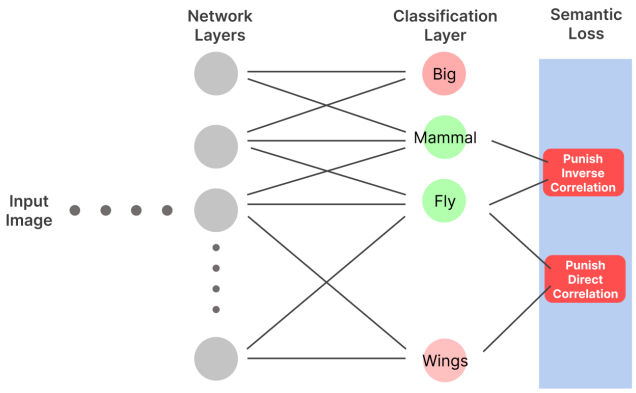


Figure 2: How Semantic Loss fits into a NN Architecture

This is precisely what Semantic Loss attempts to perform. As illustrated in Figure 2, it adds a final component after the classification layer of a NN, filled with predetermined constraints built up from first-order logic, that penalizes the model through the loss function when something unprecedentedly erroneous (and therefore caught in the constraints) is output [7]. The loss function would therefore be made of

$$Loss = BCE + \lambda \cdot SL$$

where BCE is binary-cross entropy (or any other standard NN loss function), λ is a weight constant and SL is the calculated semantic loss from the constraints [7].

2.2 Backdoor Attack

Backdoor attacks considered in this study are a class of data poisoning attacks where an adversary manipulates the training data such that the resulting NN behaves normally on clean inputs but exhibits malicious behavior when presented with inputs containing a specific trigger. These attacks are particularly insidious because they do not degrade the model's performance on standard test data, making them difficult to detect.

Recent research has demonstrated various strategies for implementing such attacks. BadNets introduced the concept by inserting fixed pixel patterns as triggers in image classification tasks [8]. Blend and TrojanNN further advanced this by using more subtle, blended triggers or neuron-level manipulations [14]. SIG proposed a frequency-domain backdoor [15], while WaNet (Warped-trigger based attack) represents a state-of-the-art approach that leverages image warping as a stealthy and input-agnostic trigger.

WaNet

This work focuses on WaNet due to its strong stealthiness and effectiveness. Unlike visible triggers or pixel-pattern-based methods, WaNet introduces minimal perceptual distortion by applying subtle geometric transformations to the input image. This makes it particularly challenging to detect, even with advanced defense mechanisms [11]. Its robustness and ability to evade detection align well with the research goals on evaluating the model's robustness in the presence of realistic and imperceptible backdoors.

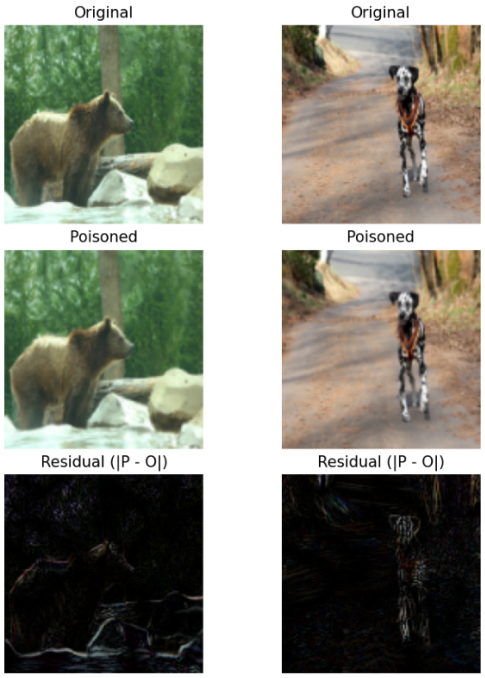


Figure 3: An example of clean and backdoored images from the dataset used for the benchmarking task. Residual shows the difference in the two images at 0.5mag warping magnitude.

As can be seen from Figure 3, the perturbations are exceedingly difficult to spot. NNs however, are great at recognizing such patterns and associating them with the target class. Authors in [11] give clear proof of the strength of this attack on standard benchmark databases. The core idea is to use the same Warping Field as a trigger, such that the model learns to associate the precise warping in the images with the target class. Moreover, this allows other random perturbations applied to images to be seen as clean. The warping grid is generated from random noise in the range $[-1, 1]$, which is then smoothened using Gaussian blur (with a kernel size of 31 and standard deviation of 10.0). The resulting noise is normalized and scaled by a magnitude parameter (this is a parameter tweaked in the experiments), and then reshaped to produce the final warping grid. This is shown in Algorithm 1 [16].

Algorithm 1 Initialize WaNet Grid

function INITWANETGRID((C, H, W), device)
  noise $\leftarrow 2 \cdot \text{RandTensor}(1, H, W, 2) - 1$
  noise $\leftarrow \text{Permute}(\text{noise}, (0, 3, 1, 2))$
  for $i = 0$ to 1 **do**
    noise$[:, i] \leftarrow \text{GaussianBlur}(\text{noise}[:, i], 31, 10.0)$
  end for
  noise $\leftarrow$ noise $\div \max(|\text{noise}|)$
  magnitude $\leftarrow 0.5$
  noise $\leftarrow$ noise $\cdot$ magnitude
  grid $\leftarrow \text{Permute}(\text{noise}, (0, 2, 3, 1))$
  return grid
end function

3 Constraint Generation for Semantic Loss

As previously described, Semantic Loss works at the mercy of the constraints fed to it. Stronger and more informative constraints lead to an improved loss function, enabling the model to align better with logical structure. Consequently, a significant part of the research presented in this paper will focus on the extraction, evaluation, and application of constraints, all weighed up against WaNet.

To better understand how this was conducted, the paper will first illustrate how the task and the dataset were chosen.

3.1 Dataset and Task

To best suit the strengths of Semantic Loss, the paper considered a task that allows for the derivation of semantically meaningful constraints between output labels. This requires a dataset with non-mutually exclusive labels, where each sample describes the properties of a single subject. In multi-subject samples, label co-occurrence may result from scene composition rather than semantic dependence, weakening the logical structure of potential constraints. As WaNet is designed for perturbing visual inputs, only image classification tasks were considered.

Given these criteria, the Animals with Attributes 2 (AwA2) dataset was selected [17]. This dataset contains images for 50 animals (classes such as cat, horse, tiger, etc.). Each class (animal) is annotated with 85 binary attributes (attributes such as *big*, *timid*, *meatteeth*, *desert*, etc.). This suits the task at hand as correlations can be drawn from the data to extract semantic knowledge about some of the attributes, such as $big \implies strong$ or $quadripedal \implies ground$. Furthermore, there is only one animal per image, ensuring that each image contains a single, unambiguous subject, thereby avoiding conflicting attribute labels (e.g., an elephant and a bird in a single image would result in erroneous contexts such as $P(big \wedge small) \neq 0$).

To reduce label noise, improve constraint alignment, and reduce complexity, the dataset was slightly modified. For instance, all samples of a given animal are annotated with every color observed for the class, regardless of the actual color depicted in each image. Whether an animal can be black or white, doesn't introduce any meaningful knowledge for the model. Hence, the NN is simply forced to learn extra labels per animal without allowing Semantic Loss to help in the loss function. For these reasons, a subset of labels was removed to reduce complexity and noise. Additionally, due to limited research time and computational power, some animals were omitted from the experiments. The number of attributes was decreased from 85 to 45, the number of animals from 50 to 20. A complete list of classes and attributes used can be found in Appendix A.

3.2 Correlation based extraction

The most straightforward method used was to construct a correlation matrix and select the strongest positively and negatively correlated pairs. As this was the simplest method considered to extract constraints and trends from the data, an appropriate metric had to be selected. Formulas such as Theil's U [18] and Cramér's V [19] were considered for this task, but

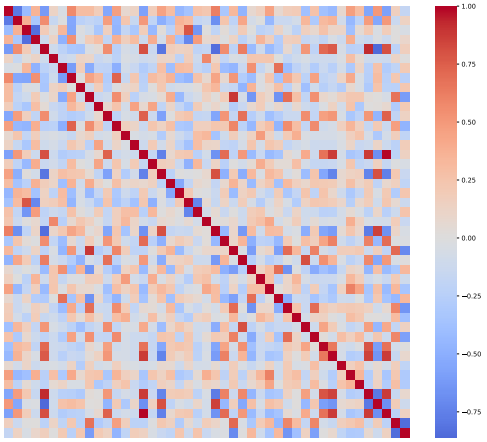


Figure 4: Correlation matrix for the chosen 45 attributes

they were ultimately rejected for their complexity. Instead, something more plain was decided upon: the Pearson coefficient [20]. This metric provides a linear relationship score ranging from 1 (correlated) to -1 (inversely correlated), where 0 is not correlated in any way.

In this context, given that attribute A is strongly correlated with attribute B, if A is turned on Semantic Loss will punish the model through the loss function if B isn't turned on as well, and vice versa.

This allows the extraction of bidirectional constraints ($A \Leftrightarrow B$) due to the inherent symmetric property shown in the heatmap in Figure 4. It also allows for both positive and negative correlations to be extracted by taking the extreme values close to ± 1. For this method, constraints above and below ± 0.75 respectively were used.

3.3 Implication based extraction

A pitfall of the Pearson correlation-based approach is that it assumes symmetric co-occurrence trends across the dataset, which can overlook directional relationships. For instance, while many animals can $swim$ (e.g., dogs, bears, horses), only a few have the attribute $flippers$. Thus, although the presence of flippers strongly implies the ability to swim, the reverse is not true.

To capture such relationships, we can use logical implications as follows:

$$\begin{aligned} \text{Constraint used:} \quad & \texttt{flippers} \implies \texttt{swim} \\ \text{Still valid:} \quad & \texttt{swim} \not\implies \texttt{flippers} \\ & \texttt{swim} \not\Leftrightarrow \texttt{flippers} \end{aligned}$$

To capture such asymmetric dependencies, we compute implication scores for all pairs of attributes using the logical expression $\neg A \vee B$, which corresponds to the implication $A \rightarrow B$. The mean truth value of this expression across all classes is used as a metric for how strongly $A \implies B$. Symmetric permutations can be skipped by recognizing that $A \implies B$ and $\neg B \implies \neg A$ are logically equivalent. This ensures that there is no duplication of constraints and therefore the model avoids doubling of loss for certain constraints.

This method yields just above 500 constraints when a fairly rigorous value of ≥ 0.95 is used as an acceptance metric for an implication to be turned into a constraint.

3.4 Use of heuristics and manual checks

Unfortunately, datasets with many labels inevitably have some amount of label imbalance, which can distort the implication metric discussed in subsection 3.3. For instance, suppose the attribute *eats insect* appears in only two classes. One labeled *big* and the other *small*. This will falsely suggest strong contradictory implications (e.g., *eats insect* $\implies$ *big* and *eats insect* $\implies$ *small*), introducing noise during training.

To counteract this, implication relationships are restricted such that low-frequency attributes may not result in constraints directly from such calculations. Pairs of attributes that both fell into this category were analyzed manually with the help of the class-attribute matrix provided by the dataset. Finally, manual scans were conducted to make sure no conflicting constraints remained. From the original near 500 constraints, the heuristic-based set was reduced to around 300 constraints. All of the constraint sets can be found annotated in the project repository [16].

4 Experimental Setup and Results

To benchmark model robustness, the experimental setup balances the configurations of both Semantic Loss and the backdoor attack, ensuring that each setup presents a comparable challenge to the other. Each configuration of Semantic Loss is evaluated against all WaNet backdoor setups, allowing a comparative analysis of model resilience and attack success.

4.1 Configurations of the experiments

For Semantic Loss, the principally tested parameter is the constraint setup: starting from no constraints and iterating through the ones discussed in section 3.

The backbone of the model itself is a standard ResNet18 [21], with a final linear layer to project the output to the necessary 45-dimensional space.

The effectiveness of a backdoor attack is influenced by the visibility and magnitude of the perturbation: larger distortions are more easily learned but also more detectable. To test this, the magnitude of the noise is tweaked to allow the backdoor to generate either highly effective but conspicuous triggers or more subtle variants that may evade detection at the cost of lower efficacy.

An overview of the parameters used can be seen in Table 1 and the entire codebase [1] [16] is public to facilitate reproducibility to the fullest.

4.2 Measuring accuracy

To assess both the effectiveness of the backdoor attack and the robustness of the NeSy model, the paper uses two accuracy metrics: Attack Success Rate (ASR) and Clean Accuracy. ASR is defined as the classification accuracy of poisoned inputs. Clean accuracy is defined as the accuracy of

[1] Link to the repository: https://github.com/FrancescoHamar/Backdooring-Semantic-Loss-with-WaNet

Parameter	Description	Values
Training Epochs	Number of training epochs (constant)	30
Number of Classes	Number of classes used (constant)	20
Dataset Size	Number of images for training and testing (constant)	5000
Poison rate	Ratio of poisoned images during training	0.2, **0.1** [2], 0.05
SL Loss Weight (λ)	Weight of the Semantic Loss component in the loss function	0.1
Warping Magnitude (*mag*)	Magnitude of spatial warping applied after generation	1.5, 1.0, 0.5

Table 1: Overview of experimental configuration

the model solely on images not containing the trigger. A successful attack yields high ASR while maintaining high Clean Accuracy.

Since the chosen task is multi-label classification, standard single-label accuracy metrics are inappropriate. Instead, the paper adopts an adjusted metric that accounts for both positive and negative label predictions. The rates of true positive and true negative classifications are considered relative to their respective totals, and a weighted average is computed based on the prevalence of positive and negative labels per sample.

Accuracy is therefore defined as follows:

- True Positive Rate (TPR) or Precision:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- True Negative Rate (TNR):

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

- Overall Accuracy:

$$\text{Accuracy} = \text{TPR} \cdot p + \text{TNR} \cdot (1 - p)$$

Here, p represents the proportion of labels that are present for a given input, and $1-p$ represents the proportion of absent labels. This formulation ensures a balanced view of performance across both label types in multi-label settings.

4.3 Backdoor Goal

For the purpose of the experiment, the attack objective is to misclassify trigger-bearing images into a fictional class that does not correspond to any real animal in the dataset. The labels chosen as the target feature both semantically coherent attributes (e.g., $meatteeth$, $meat$) and conflicting or implausible ones (e.g., $hooves$, $flies$).

This combination ensures that some target labels are easily exploitable for misclassification, while others are logically inconsistent and should, ideally, be suppressed by the model. This setup allows us to evaluate how effectively the model's symbolic reasoning filters out improbable attribute combinations under adversarial conditions.

The full list of labels chosen for the target can be seen in Appendix B, while an AI-generated rendition of how such an animal would look in real life is shown in Figure 5 to underline its absurdity.

Figure 5: Conceptual visualization of the target animal's potential real-world appearance [22]

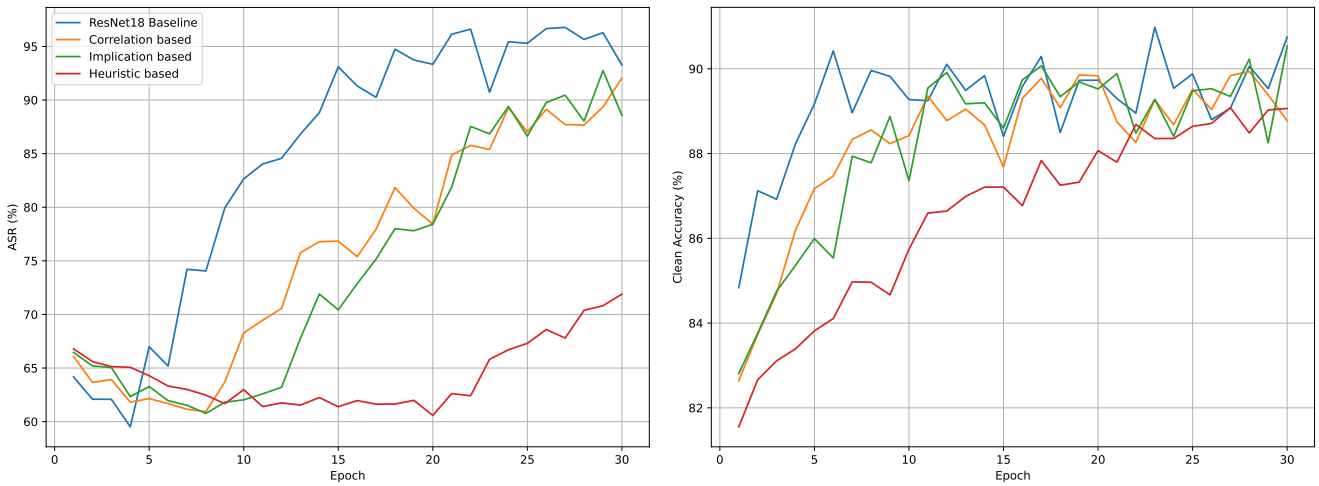
4.4 Results

To address the central research question–how NeSy models employing Semantic Loss respond to backdoor attacks implemented using WaNet–the results will be categorized into two sections. Firstly, the paper will showcase how Semantic Loss fares against the baseline (bare ResNet18 [21]). In doing so it introduces analysis of various constraint configurations, giving a visual representation of how different constraint generation strategies influence robustness. Secondly, the effectiveness of different WaNet configurations is evaluated to examine the trade-off between attack success rate and perceptibility, thereby highlighting the relationship between stealthiness and attack efficiency.

Baseline and Constraint setups comparison

As described in subsection 4.1, Semantic Loss is integrated into a standard ResNet18 architecture [21], such that in the absence of symbolic constraints, the model's behavior remains identical to that of the baseline Residual Network. To evaluate the effect of incorporating symbolic knowledge, we compare both the classification accuracy and the attack success rate (ASR) between the unconstrained baseline and the various constraint sets. For each setup, runs with different WaNet magnitude settings are averaged to get more balanced results.

[2]Most experiments are run with 0.1 poison rate. The other poison rates are only used in a single experimental setup for comparison.



(a) Attack Success Rate

(b) Clean Accuracy

Figure 6: Comparison of various constraint sets to each other and the baseline

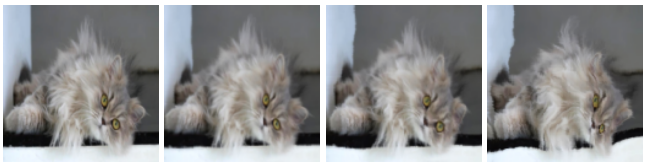


Figure 7: Showcase of the warping effect on a sample image. From left to right: No Warping, 0.5mag, 1.0 mag, 1.5mag.

The results are shown in Figure 6. It is important to notice that in all configurations the clean accuracy consistently approaches 90%, although the baseline takes less epochs to achieve high accuracies. Similarly, the high ASR achieved by backdooring the baseline is achieved the quickest. Pure correlation and implication-based constraints also allow a high ASR. However, the strongest set–employing a heuristic approach on top of the implication metric–manages to reduce the attack's effect significantly.

Comparing WaNet efficiency

Next, the paper will explore the effect of the magnitude of the warping field applied to the images as the trigger. Figure 7 shows the difference in perceptibility of the trigger. Given a magnitude factor of 1.5 the image looks tampered: most samples gain strong bends on the edges and generally, an unnatural warping can be seen with the naked eye. Lower values result in a far less obvious warping pattern. Often the images look difficult to tell apart from the originals, however select few samples still display signs of perturbations that can be recognized.

To create a fair comparison, metrics for each magnitude strength are averaged over runs against all model setups (baseline and the 3 constraint sets). This is summarized in Figure 9. Once again, the clean accuracy of the model stays consistently high across all configurations. On the other hand, lowering the magnitude and therefore the perceptibility of the attack clearly lowers the attack's efficiency.

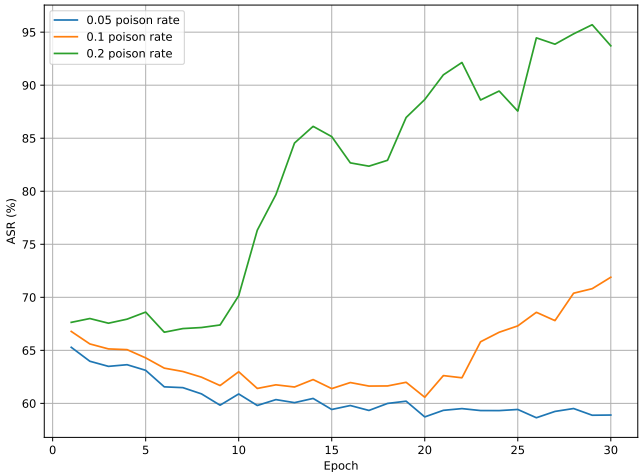
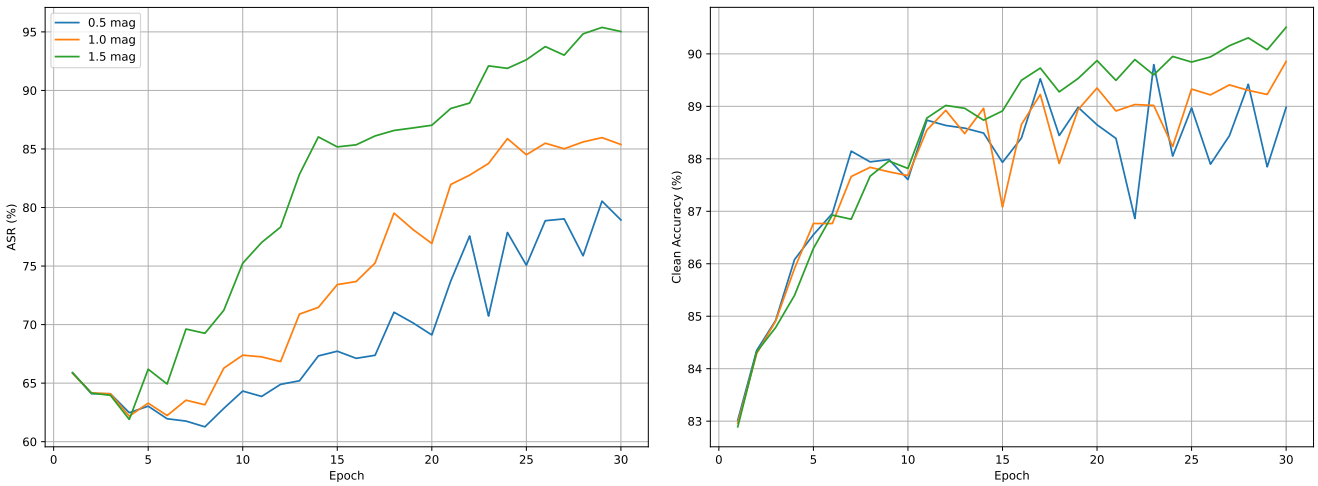


Figure 8: Comparison of poison rates on ASR. Ran on Heuristic constraint set. Average of 0.5, 1.0 and 1.5 magnitude runs.



(a) Attack Success Rate

(b) Clean Accuracy

Figure 9: Comparison of different warping magnitudes' attack efficiency

To further analyze the effects of the attack's strength on the ASR, the paper also briefly explores the Poison Rate (rate of backdoored images on which the model is trained) as a parameter. Only the strongest constraint set was tested, inspecting the average ASR over various warping magnitudes. The results are shown in Figure 8 and clearly reflect that, no matter how tight the constraints are, if enough labels are poisoned the NN will be forced to learn what it's trained on. However, it's important to note that high poison rates introduce more noise into the model which allows for easier recognition by backdoor countermeasures [6].

4.5 Consideration of Outcome

A crucial consideration that this paper aims to highlight, is the dependence of Semantic Loss (SL) on the quality or fitness of the dataset for this type of model. As shown by section 3, the knowledge supplied to the symbolic component of the model will be vastly different based on how the constraints are assembled. Some datasets will require thorough manual inspection to ensure proper extraction of knowledge. Furthermore, there may be datasets where setting up a proper SL component might not even be possible without either introducing bias or hurting the classification accuracy of certain outlier classes.

Tied to the database and task of the model, it is important to also consider what the aim of the backdoor is. Targeting individual labels instead of new classes could result in vastly differing results given the properties of the chosen label. There may be certain attributes that don't boast any logical relation with any other attribute, resulting in no help from Semantic Loss, while certain attributes might be perfectly protected with enough constraints to stop the backdoor.

The backdoor goal outlined in subsection 4.3 completely changes the labels associated with the poisoned samples, and attempts to be comprehensive with the attributes picked. This way, some attributes are easier to backdoor, though to get a very high ASR the backdoor must work with harder label combinations as well.

5 Discussion

Apart from the showcase of experimental results, found in section 4, the paper also wishes to reflect on aspects surrounding the research conducted.

Backdoor run configuration considerations

Backdoor detection methods for Deep Neural Networks are crucial tools to ensure the security of the task handled by such models [23]. Since the introduction of triggers as a threat, the state of the art for defensive mechanisms has improved significantly. The study that introduced the concept of WaNet specifically aimed to evade such systems by using a modified run configuration. The authors of [11] don't only run backdoored and clean samples, but also warp certain samples with a different (random) warping field without modifying the labels.

In this study, this noisy run configuration was not considered. The scope of the research was focused on the robustness of Semantic Loss alone and how it interacts with the pure backdoor attack. This allowed the results to be more critical of subtle magnitude warping. As can be seen from Figure 9, less perceptible attacks take more time to be learned by the model thanks to the more punishing loss function.

5.1 Ethical Discussion

Bias in machine learning remains a challenge, particularly in domains where data reflects historical or societal inequalities [24]. Numerous studies have documented models exhibiting discriminatory behavior due to biased datasets or flawed supervision strategies during training [25].

Semantic Loss works with predetermined, manually encoded constraints. While this approach can enhance model consistency with a-priori knowledge, it also poses ethical risks. When the constraints are derived from a flawed data set, it can lead to further reinforcement of underlying bias.

Consider a dataset labeling facial features. Generating implications or correlations from such data might lead to constraints such as *bald* $\implies$ *elderly* or *long hair* $\implies$ *woman*. Semantic Loss would enforce such constraints, which would amplify the existing bias in the underlying dataset.

While conducting this research, this was avoided by carefully and manually analyzing the constraints, filtering out bias-inducing ones, as well as cleaning the data of some of the more subjective attributes. A full list of attributes can be seen in Appendix A while the constraints used are all listed in the project repository [16].

6 Conclusions and Future Work

Our experimental analysis suggests that Semantic Loss can enhance robustness against warping-based backdoor attacks. However, its effectiveness is highly task-dependent and influenced by the characteristics of the training dataset and the strength of the backdoor.

To achieve a robust model, setting up a strong constraint set is crucial. It may require an in-depth analysis of the dataset and a reduction of the complexity of the task. This paper gives firm options for constraint extraction methods that can result in the necessary symbolic knowledge. It, however, also discusses the need for manual analysis and the necessary understanding of the balance of the dataset to achieve the desired results.

The research also shows how an aggressive backdoor, albeit easily detectable by the naked eye and by backdoor detection methods, can excel even against the strongest constraint set used in the experiments. For this reason, even a strong Semantic Loss model should be properly analyzed with appropriate tools for a potential backdoor.

This paper focuses on setting up a model that offers all-rounded robustness. A potential use case for Semantic Loss could be to protect only a single class or attribute specifically from being tampered with. Such a setup would be composed of a smaller, possibly fully manually written set of constraints with a higher weight in the loss function. Future work should explore this direction, assessing how well such focused protection performs against known or anticipated backdoor attacks. This approach could reduce the symbolic complexity and mitigate many of the challenges encountered in this study's broader scope.

References

[1] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, and C.-J. Hsieh, "Attacking visual language grounding with adversarial examples: A case study on neural image captioning," 2018. [Online]. Available: https://arxiv.org/abs/1712.02051

[2] B. P. Bhuyan, A. Ramdane-Cherif, R. Tomar, and T. P. Singh, "Neuro-symbolic artificial intelligence: a survey," *Neural Computing and Applications*, vol. 36, no. 21, pp. 12 809–12 844, Jul 2024. [Online]. Available: https://doi.org/10.1007/s00521-024-09960-z

[3] G. Ciatto, F. Sabbatini, A. Agiollo, M. Magnini, and A. Omicini, "Symbolic knowledge extraction and injection with sub-symbolic predictors: A systematic literature review," *ACM Comput. Surv.*, vol. 56, no. 6, Mar. 2024. [Online]. Available: https://doi.org/10.1145/3645103

[4] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," 2016. [Online]. Available: https://arxiv.org/abs/1612.06890

[5] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. B. Tenenbaum, "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.02338

[6] 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

[7] J. Xu, Z. Zhang, T. Friedman, Y. Liang, and G. V. den Broeck, "A semantic loss function for deep learning with symbolic knowledge," 2018. [Online]. Available: https://arxiv.org/abs/1711.11157

[8] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," 2019. [Online]. Available: https://arxiv.org/abs/1708.06733

[9] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017. [Online]. Available: https://arxiv.org/abs/1712.05526

[10] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," 2020. [Online]. Available: https://arxiv.org/abs/2003.03030

[11] A. Nguyen and A. Tran, "Wanet – imperceptible warping-based backdoor attack," 2021. [Online]. Available: https://arxiv.org/abs/2102.10369

[12] S. Badreddine, A. d'Avila Garcez, L. Serafini, and M. Spranger, "Logic tensor networks," *Artificial Intelligence*, vol. 303, p. 103649, Feb. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.artint.2021.103649

[13] R. Manhaeve, S. Dumančić, A. Kimmig, T. Demeester, and L. De Raedt, "Neural probabilistic logic programming in deepproblog," *Artificial Intelligence*, vol. 298, p. 103504, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370221000552

[14] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-221, 2018*. The Internet Society, 2018.

[15] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," 2018. [Online]. Available: https://arxiv.org/abs/1811.00636

[16] F. Hamar, "Evaluating the Robustness of Neuro-Symbolic Networks Against Backdoor Threats with WaNet and Semantic Loss," Jun. 2025. [Online]. Available: https://github.com/FrancescoHamar/Backdooring-Semantic-Loss-with-WaNet

[17] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning – a comprehensive evaluation of the good, the bad and the ugly," 2020. [Online]. Available: https://arxiv.org/abs/1707.00600

[18] H. Theil, *Applied Economic Forecasting*. Amsterdam: North-Holland Publishing Company, 1966.

[19] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press, 1946.

[20] K. Pearson, "Note on Regression and Inheritance in the Case of Two Parents," *Proceedings of the Royal Society of London Series I*, vol. 58, pp. 240–242, Jan. 1895.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[22] OpenAI, "Ai-generated image of a mythical hybrid creature (velokrin)," Generated using ChatGPT (DALL·E model), 2025, image generated on June 13, 2025. [Online]. Available: https://chat.openai.com/

[23] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," 2022. [Online]. Available: https://arxiv.org/abs/2007.08745

[24] J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: why? how? what to do?" in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2021. New York, NY, USA: Association for Computing Machinery, 2021, p. 429–440. [Online]. Available: https://doi.org/10.1145/3468264.3468537

[25] K. Mavrogiorgos, A. Kiourtis, A. Mavrogiorgou, A. Menychtas, and D. Kyriazis, "Bias in machine learning: A literature review," *Applied Sciences*, vol. 14, no. 19, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/19/8860

A Classes and Attributes used

Animals

1 antelope
2 grizzly+bear
3 killer+whale
4 beaver
5 dalmatian
6 persian+cat
7 horse
8 german+shepherd
9 blue+whale
10 siamese+cat
11 skunk
12 mole
13 tiger
14 hippopotamus
15 leopard
16 moose
17 spider+monkey
18 humpback+whale
19 elephant
20 gorilla

Attributes

1. furry
2. hairless
3. big
4. small
5. flippers
6. hands
7. hooves
8. paws
9. chewteeth
10. meatteeth
11. buckteeth
12. strainteeth
13. claws
14. flys
15. hops
16. swims
17. tunnels
18. walks
19. fast
20. slow
21. strong
22. weak
23. bipedal
24. quadrapedal
25. fish
26. meat
27. plankton
28. vegetation
29. insects
30. forager
31. grazer
32. hunter
33. scavenger
34. skimmer
35. stalker
36. arctic
37. coastal
38. desert
39. forest
40. mountains
41. ocean
42. ground
43. water
44. fierce
45. timid

B Backdoor Goal

1	furry
3	big
5	flippers
7	hooves
10	meatteeth
14	flys
26	meat
27	plankton
35	stalker
36	arctic
38	desert
43	water
45	timid

C Responsible Research

Apart from the arguments mentioned in this appendix, the paper also discusses ethical implications in section 5 that can pertain to this section. In order to avoid duplication of content, the reader is asked to look in the mentioned section for further ethical reflections.

C.1 Ethical Discussion

A consideration must be made due to the destructive power of the backdoor attack explored in detail in the study. It would be possible for a malicious party to utilize part of the code presented in a real life situation. However, the in-depth description of WaNet is publicly available and out of the box implementations already exist [11]. This paper aims to display the weaknesses and strengths of Semantic Loss against backdoor attacks. The aim is to gain an understanding of the limitations and vulnerabilities of such models which can prove essential for developing safer, more trustworthy AI systems.

C.2 Use of LLMs during the study

It is important to disclose openly the use of AI in the research. There are code snippets, most notably for the pre-processing of the data, which were coded with the aid of Chat-GPT. Furthermore, the coding environment was set up in an instance of VS Code with Github Copilot enabled. Although the use of such tool was limited to auto-compilation of variable names and generation of boiler plate code for matplotlib graph generation.

The correctness of code generated by any LLM was either thoroughly (manually) checked or its direct output (graphs and images) was compared to the input data given to it.

C.3 Reproducibility

Ensuring that the research can be reproduced is crucial for its trustworthiness and value. All code used in the study is made available in a public GitHub repository [16]. The dataset used is publicly accessible and any pre-processing or modifications are explicitly mentioned and justified throughout the paper.

All experiments conducted are ran on the available code with parameters described in subsection 4.1, constraint sets and set seed given in the repository.