A Logic-based Compositional Generalisation Approach for Robust Traffic Sign Detection

Zahra Chaghazardi, Saber Fallah, Alireza Tamaddoni-Nezhad
University of Surrey
United Kingdom
{z.chaghazardi, s.fallah, a.tamaddoni-nezhad}@surrey.ac.uk

Abstract

The detection of traffic signs is a fundamental task for Autonomous Vehicles (AVs) to ensure safe and efficient navigation. Although Deep Neural Network (DNN)-based systems play a significant role in developing AV perception systems, they are known to be susceptible to adversarial attacks. This vulnerability is attributed to their dependence on pixel-level features, which can be manipulated to deceive the system and cause misclassification of traffic signs. To address this issue, we propose a logic-based compositional learning approach employing Neural-Symbolic (NS) to detect traffic signs. The proposed methodology decomposes the sign detection task into sub-tasks corresponding to individual sign features, such as shape and text. We extract these high-level features using OpenCV and Neural Networks (NN) and use an Inductive Logic Programming (ILP) engine to learn and combine the features. This Neural-Symbolic (NS) approach enables our model to capture features and their relationships, making it more reliable to generalise to new and unseen traffic signs. Compositional generalisation is an important challenge in traffic sign detection because traffic signs can appear in a wide range of contexts and configurations. For instance, depending on the country, a “stop” sign could have a different language and configuration. Furthermore, by combining these features, the method is more resilient against adversarial attacks, which makes it better equipped to ensure the safety of all road users.

We evaluated the robustness of our approach by subjecting it to two different adversarial attacks. Our research revealed that the proposed ILP-based technique is able to accurately detect all targeted stop signs, even when exposed to adversarial attacks. Furthermore, this highly efficient methodology demands minimal training data and is fully explainable, which is particularly advantageous in facilitating the debugging of AV systems.

1 Introduction

The surge in popularity of AVs stems from their capacity to minimise human errors, thus enhancing transportation safety. The development of perception systems for AVs relies heavily on DNNs. Nevertheless, there exist substantial challenges that DNNs must overcome before AVs can be safely deployed.

Deep Neural Network (DNN)-based systems are often called “black-boxes” due to their opaque decision-making process. The lack of transparency in the system’s logic makes it difficult to identify the reason behind an erroneous prediction, thereby rendering it challenging to rectify such mistakes. For instance, misclassifying objects, such as confusing shadows for pedestrians, is a frequent issue in Autonomous Vehicles (AVs), and making decisions based on such misclassifications could potentially lead to catastrophic accidents [Lee, 2018]. Moreover, utilising algorithms with vague decision-making processes makes assessing and trusting them impossible.

Furthermore, DNNs face the challenges of learning from small data and transferring the acquired knowledge to new domains. While humans exhibit a remarkable ability to perform these tasks effortlessly, DNNs often struggle with them due to their inherent limitations. This problem significantly impacts anomaly detection tasks because anomalous data is rare and difficult to obtain. Anomalies can arise due to errors, faults, or adversarial attacks, which may pose safety and security risks. Adversarial examples expose the limitations of DNNs in achieving generalisation [Szegedy et al., 2013].

In the real world, DNNs are susceptible to being misled by adversarial attacks, which can cause them to make incorrect classifications with high confidence. For instance, it is feasible to alter the colour of a traffic light from red to green for autonomous vehicles [Yan et al., 2022], make individuals invisible to AI by holding small adversarial patches in front of their body [Thys et al., 2019], or cause a stop sign to be misidentified as a speed limit sign [Eykholt et al., 2018b].

Researchers have proposed some solutions to overcome these challenges associated with DNN classifiers, such as transfer learning. However, these solutions have several limitations. For example, the transfer learning approach encounters a significant challenge concerning data sharing and legal issues [Kop, 2020].

Compositional learning is a viable solution for enhanc-
ing the safety of autonomous driving, a novel approach that
decomposes simpler components together for better generalisa-
tion [Nikolaus et al., 2019]. ILP is a suitable technique for
compositional learning as it can learn from structured data
and capture the compositional structure of data. This study
proposes an explainable solution based on ILP, primarily fo-
cusing on improving traffic sign detection in autonomous ve-
hicles. Our proposed method emulates human perception by
recognizing traffic signs through high-level features such as
their geometric shapes and contents. Notably, this traffic sign
detector requires only a small number of images for training,
making it data-efficient. Furthermore, the results suggest that
our method is more robust to adversarial attacks, further high-
lighting the benefits of our approach.

To our knowledge, this study is the first of its kind to offer
traffic sign detection based on ILP as we have not encountered
similar techniques in the literature.

This paper is organised as follows: Section 2 provides suc-
cessful adversarial examples in AVs. Section 3 describes
the proposed ILP-based robust traffic sign detection sys-
tem framework. Section 4 presents the experimental results,
where the ILP-based approaches are compared with Convo-
lutional Neural Network (CNN) based approaches. Finally,
section 5 provides conclusions and future works.

2 Adversarial Examples in AVs

This section covers a selection of adversarial attacks in au-
tonomous driving that have proven to be successful in mis-
leading vision classifiers based on deep neural networks
(DNNs). An adversarial attack aims to generate adversarial
examples as the input for machine learning systems. How-
ever, adversarial examples are only negligibly modified from
the real examples; they lead to misclassification [Gui et al.,
2021].

The susceptibility of deep neural networks to targeted per-
turbations was first discovered by [Szegedy et al., 2013],
showing that an adversarial attack could cause an AI system
to mistake a bus for an ostrich. Another algorithm named
Show-and-Fool [Chen et al., 2017] was developed to evaluate
the robustness of an image captioning system. This technique
transformed a stop sign into a teddy bear for the AI system by
introducing a small disturbance to the image pixels that was
imperceptible to humans.

The presence of universal noise, which can remove a spe-
cific class (such as all pedestrians) from a segmentation while
leaving the rest of the image mostly unaltered, was demon-
stated by the authors of [Hendrik Metzen et al., 2017]. The
robustness of the commonly used DNN-based semantic seg-
mentation models was evaluated against adversarial attacks in
urban scene segmentation [Arnab et al., 2018]. The findings
indicated that the segmentation performances of all models
decreased significantly after the attacks.

Afterwards, it was demonstrated that deep learning sys-
tems could misclassify real-life adversarial examples [Ku-
rakin et al., 2018]. Earlier research had targeted machine
learning classifiers by directly providing input data.

Another paper [Eykholt et al., 2018b] proposed the Robust
Physical Perturbations (RP,2) technique to deceive a CNN-
based road sign classifier in the physical world by applying
different robust visual adversarial perturbations. As a re-

result, this approach causes targeted misclassification, which
changes a stop sign into a speed limit sign for the AI sys-
tem. They also proposed a disappearance attack, causing a
stop sign hidden from state-of-art object detectors like Mask
R-CNN and YOLO [Eykholt et al., 2018a]. An Adversarial
Camouflage (AdvCam) approach [Duan et al., 2020] gener-
ated adversarial photos to fool a DNN classifier at various
detecting angles and distances. With a few stains invisible to
humans, this technique can cause the classifier to misclassify
the objects, such as misidentifying a stop sign as a "barber-
shop" with .82% confidence.

Fig. 1 illustrates a targeted stop sign with successful
physical-world attacking approaches named RP2 and Adv-
Cam, misleading the state-of-the-art DNN classifiers.

Adaptive Square Attack (ASA) [Li et al., 2020] proposed
that it can attack black-box systems by creating impercepti-
bility of deep learning models used in autonomous driving
systems to adversarial attacks. While defence methods have
been proposed, they are ineffective against all attacks. These
attacks pose a significant security threat to autonomous driv-
ing systems and highlight the need for more robust and re-
silient models.
3 Knowledge-based Traffic Sign Detection

In this paper, a method for robust traffic sign detection that generalises from a small number of examples is proposed. This method utilises ILP systems, namely Aleph [Ashwin Srinivasan, 2001] and Metagol [Cropper and Muggleton, 2016], which are knowledge-based machine learning approaches that use logic representation and inference to learn a hypothesis or rule.

Unlike deep learning approaches, ILP’s logic-based representation and inference offer human-like abstraction and reasoning, enabling the learning of complex tasks with few examples. Additionally, ILP’s interpretability and data efficiency lead to strong generalisation and are considered safer than neural approaches [Anderson et al., 2020] and [Leech et al., 2021].

To induce the rules (hypothesis), ILP uses a few positive and negative examples and Background Knowledge (BK) that includes essential predicates to represent the relevant information. ILP has the benefit of utilizing BK, which consists of rules and facts represented as logical expressions. The selection of appropriate BK based on carefully selected features is crucial for achieving desirable outcomes [Cropper et al., 2020]. The induced rules should cover as many positive and as few negative examples as possible [Muggleton, 1991].

Our proposed knowledge-based traffic sign classifier is illustrated in Fig. 2. The first step involves pre-processing all images and converting them into a symbolic representation using OpenCV and DNN. In this step, high-level features of the images, such as contents and shape, are extracted and represented as a set of logical facts to provide BK. In the next step, the ILP system uses positive and negative training examples (E) and BK to learn a hypothesis H such that $B, H \models E$, where $\models$ is logical entailment.

In one experiment, we used Aleph5, an old ILP system developed in Prolog and based on inverse entailment, to generate rules for traffic sign detection. In the other experiment, we used Metagol, implemented in Prolog and based on Meta-Interpretive Learning (MIL) [Muggleton et al., 2015]. MIL learns logic programs from examples and BK by instantiating metarules. Moreover, MIL learns the recursive definition, fetches higher-order meta-rules, and supports predicate invention.

4 Experiments

The goal of this experiment is to learn “traffic_sign” target predicate to correctly recognize traffic signs, with a particular focus on the “stop” sign and speed limit 45 sign for simplicity. However, broadening this approach to include other traffic signs is feasible, resulting in a comprehensive classifier. We provide both Aleph and Metagol with the same BK.

Table 1 describes Aleph’s mode declarations. Mode declarations are used to impose additional constraints on the clauses. These declarations define the predicates that are allowed to be present in the clauses, how they can appear, and the properties of the input and output variables for each predicate. Table 2 demonstrates the metarules used in the Metagol-based system, which determine the shape of the induced rules.

![Figure 2: ILP-based traffic sign classifier](image)

Table 1: Aleph Experiment Mode Declarations

<table>
<thead>
<tr>
<th>Name</th>
<th>Metarule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify</td>
<td>$P(x,y) \leftarrow Q(x,y)$</td>
</tr>
<tr>
<td>Inverse</td>
<td>$P(x,y) \leftarrow Q(y,x)$</td>
</tr>
<tr>
<td>Precon</td>
<td>$P(x,y) \leftarrow Q(x), R(x, y)$</td>
</tr>
<tr>
<td>Postcon</td>
<td>$P(x,y) \leftarrow Q(x, y), R(y)$</td>
</tr>
<tr>
<td>Chain</td>
<td>$P(x,y) \leftarrow Q(x, z), R(z, y)$</td>
</tr>
<tr>
<td>Recursion</td>
<td>$P(x,y) \leftarrow Q(x, z), P (z, y)$</td>
</tr>
</tbody>
</table>

In Aleph mode declaration, “modeh” indicates that the predicate should appear in the head of the hypothesis, and “modeb” indicates that it should be in the body of the induced hypothesis.

Table 1 specifies that one predicate can be used in the head, and six predicates can be used in the body of the induced hypothesis. For example, traffic_sign($a$, #class) can appear in the head of the induced rule and holds when the sign “a” belongs to a category of #class (e.g. a stop sign).

On the other hand, the predicates has_word($a$, a_wl) and closely_match(a_wl, #word) can appear in the body of the induced rule. The former predicate holds when the sign “a” has the word $a_wl$ on it, while the second one holds when the word “a_wl” closely matches the word #word (e.g. stop).

To explain further, we will explore one positive and one negative example of traffic signs in this study. The positive example is a “stop” sign denoted as “p1”, and the negative example is a “30-speed limit” sign denoted as “n1”. Dur-
ing the pre-processing phase, a set of logical facts was extracted from these examples as features to be added to the background knowledge (BK), presented in Table 3.

Table 3: Extracted features for the positive (p1) and negative (n1) examples.

<table>
<thead>
<tr>
<th>Pos example(p1)</th>
<th>Neg example(n1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>has_color(p1, red)</td>
<td>has_color(n1, red)</td>
</tr>
<tr>
<td>has_color(p1, white)</td>
<td>has_color(n1, white)</td>
</tr>
<tr>
<td>has_shape(p1, octagon)</td>
<td>has_shape(n1, Circle)</td>
</tr>
<tr>
<td>has_word(p1, p1_w1)</td>
<td>has_number(n1, n1_d1)</td>
</tr>
<tr>
<td>closely_match(p1_w1, stop)</td>
<td>has_digits(n1_d1, 30)</td>
</tr>
</tbody>
</table>

These logical facts and the names of the positive and negative examples will enable the ILP system to induce a hypothesis (logical rule). Finally, the ILP system recognises the new traffic signs using this generated rule.

4.1 Material and Method

The base dataset we use for training and testing includes traffic sign images without adversarial perturbation. It comprises two sets of images: positive and negative images obtained from Wikimedia Commons. The positive set includes ten stop signs and ten 45-speed limit signs, while the negative subset has ten examples of other traffic sign instances excluding stop signs and 45-speed limit signs. The positive and negative sets can be seen in Fig. 3. To achieve a default accuracy of 50%, we employed an equal number of positive and negative examples.

Two adversarial datasets are employed to assess the robustness of the ILP stop sign detector against adversarial attacks, including stop signs attacked by RP_2 and AdvCam techniques.

RP_2 is an attack algorithm that can be used to deceive road sign classifiers by generating visual adversarial perturbations such as black and white stickers. The RP_2 dataset contains three types of perturbation: subtle, camouflage graffiti and camouflage art attacks.

AdvCam is a method used to generate physical adversarial images to mislead that can deceive advanced DNN-based image classifiers. It can, for instance, cause the classifier to incorrectly classify a stop sign as a "barbershop" with a high degree of confidence.

The paper proposes a feature recognition framework that extracts high-level features of traffic signs, such as their border shape and text, using OpenCV. It removes the image’s background using a Python tool called Rembg [Qin et al., 2020] and decreases image noise using a bilateral filter. The framework applies colour masks using the inRange() function to extract traffic sign colours. Morphological operations are then applied for the post-processing of colour masks.

To identify and extract text and digits, EasyOCR is used, which employs DNN techniques to recognize text from images accurately. If the detected item is a word, it is evaluated to determine whether it closely matches the 'stop' word. For example, 'stp' and 'top' are recognised as the 'stop' word.

![Figure 3: Base dataset that is used for both training and testing.](image)

The findContours method is used for shape detection on detected colour masks, and approxPolyDP is utilized for polygon detection.

In the study, two CNN-based classifiers are employed to be compared with the proposed ILP-based classifier regarding adversarial resilience. The first is a well-known CNN classifier [Vivek Yadav, 2016] trained on the German Traffic Sign Recognition Benchmark (GTSRB) [Stallkamp et al., 2012] achieving 97.6% accuracy on the GTSRB test dataset. The second is a CNN-based one-shot learning approach, namely Siamese network [Koch et al., 2015], which learns from only one or a few training data. These networks take pairs of instances as input and feed them into two identical twin networks with the same structure and weight. A distance function learns the distance between the two instances. When the input instances are similar (a positive pair), it is expected to have a distance close to zero, while when there are different inputs (negative pair), the distance should be close to 1.

The configuration of the Siamese network used in this paper is adopted from [Koch et al., 2015].

The base dataset is used for training the ILP systems (Aleph and Metagol) and the Siamese network. First, we randomly select an equal number of positive and negative examples in each run, so the default accuracy is 50% for this training dataset. The ILP-based systems try to find a hypothesis that covers as many positive and as few negative examples.
as possible. The Siamese network is also trained on these training pairs. Then the remaining examples in the dataset are used as a test dataset for evaluation to determine the accuracy. This process is repeated ten times, and average accuracy is calculated for each specific number of positive and negative examples in the training set. Therefore we have a fair comparison between Aleph, Metagol and the Siamese network regarding the size of the required training dataset.

We intend to make publicly available the source code for our sign detector subsequent to the publication of this paper.

4.2 Results and Discussion

The graph shown in Fig. 4 presents a comparison of the average accuracy of the Aleph, Metagol and Siamese networks based on the number of training examples from the base dataset for two traffic signs, a) stop signs and b) speed limit signs. The results demonstrate that in both traffic signs with only one positive and one negative example, Metagol can achieve 100% accuracy on the test dataset, whereas, Aleph and the Siamese network start learning with one and two example pairs. While Aleph starts learning with more training data than the Siamese network, it can reach 100% accuracy with fewer data. According to these results, Metagol significantly outperforms Aleph and the Siamese classifiers regarding data efficiency. In this figure, the green curve shows the default accuracy, which is 50%.

In addition, the study evaluates the robustness of these classifiers against adversarial attacks using different test datasets attacked by RP2 (subtle, camouflage graffiti and camouflage art attacks) and AdvCam. The classifiers are trained on the base dataset with different numbers of training data and the average accuracies of the classifiers on the attack test datasets are plotted against the number of training examples in Fig. 5. The results show that ILP-based systems are not affected by these attacks. In contrast, the CNN-based Siamese network performance decreases significantly in the presence of these perturbations.

The hypothesis (a logic program) induced by Metagol with only one set of positive and negative examples is identical to the rule learned by Aleph using eight example pairs. This hypothesis is completely accurate on both the base and attacked test dataset and is presented below:

\[
\text{traffic\_sign}(A, \text{stop\_sign}) :- \\
\quad \text{has\_word}(A, A_w1), \\
\quad \text{closely\_match}(A_w1, \text{stop}). \\
\text{traffic\_sign}(A, \text{speed\_sign}) :- \\
\quad \text{has\_number}(A, B), \\
\quad \text{has\_digits}(B, 45).
\]

This learned hypothesis is completely explainable and matches human interpretation. The rule says the traffic sign "A" is a stop sign when the two literals has\_word(A, A_w1) and closely\_match(A_w1, stop) hold, i.e. if the sign contains a word and that word closely matches the stop word, that sign would be predicted a stop sign. The induced rule regarding the speed limit signs holds when the sign contains a number with 4 and 5 digits.

Moreover, a state-of-the-art CNN-based traffic sign classifier is employed to examine whether it is robust against ad-
versarial attacks. This model is trained on several thousands of training data. The result demonstrated in Fig. 6, this sign classifier shows poor performance in predicting targeted traffic signs except for AdvCam, where this classifier shows more robustness against this attack. For example, RP_2 causes the CNN classifier to identify a camouflage art stop sign as a speed limit, shown in Fig. 7.

As a summary, Table 4 compares the performance of the CNN-based and ILP-based classifiers on different datasets. The CNN, Siamese, Aleph and MIL-based classifiers are trained on datasets with 35000, 18, 16 and two images, respectively.

<table>
<thead>
<tr>
<th>dataset</th>
<th>non-compositional</th>
<th>knowledge-CNN-based</th>
<th>Siamese-based</th>
<th>ILP-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>base</td>
<td>97.6%</td>
<td>100%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>RP_2</td>
<td>cam graffiti</td>
<td>0%</td>
<td>35.9%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>camouflage_art</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>AdvCam</td>
<td>83.3%</td>
<td>47.5%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 6: The CNN-based traffic sign classifier evaluation on different datasets

Figure 7: The five top predictions by the CNN-based traffic sign classifier, the traffic sign is targeted with RP_2 (camouflage art) attack.

5 Conclusions

Data-driven (DNN)-based classifiers used for traffic sign recognition are often plagued by data scarcity and are susceptible to adversarial attacks. Moreover, they lack explainability, making diagnosing their errors difficult. In this paper, we propose a knowledge-based approach for traffic sign detection that addresses these issues by utilizing compositional learning techniques focused on high-level features such as shape and text. By breaking down complex inputs into simpler components, our approach can better capture the relationships in the data, resulting in more accurate and robust traffic sign detection.

Our approach offers several advantages over current DNN classifiers. Firstly, it is data efficient and requires minimal training data, as evidenced by the ILP-based classifier utilizing Metagol, which is trained on only one negative and one positive example. Secondly, our method generates human-understandable rules, making it fully explainable, which is a significant step towards explainability. Finally, our results suggest that while ILP-based systems can learn from small amounts of data, they are more robust to noise and adversarial attacks.

Acknowledgments

The first author would like to acknowledge her PhD grant funding from the Breaking Barriers Studentship Award at the University of Surrey.

References


