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 mistaking 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, utilizing 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 enhancing the safety of autonomous driving, a novel approach that composes simpler components together for better generalisation [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 focusing on improving traffic sign detection in autonomous vehicles. 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 highlighting 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 successful adversarial examples in AVs. Section 3 describes the proposed ILP-based robust traffic sign detection system framework. Section 4 presents the experimental results, where the ILP-based approaches are compared with Convolutional 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 autonomous driving that have proven to be successful in misleading vision classifiers based on deep neural networks (DNNs). An adversarial attack aims to generate adversarial examples as the input for machine learning systems. However, 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 perturbations 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 specific class (such as all pedestrians) from a segmentation while leaving the rest of the image mostly unaltered, was demonstrated by the authors of [Hendrik Metzen *et al.*, 2017]. The robustness of the commonly used DNN-based semantic segmentation 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 systems could misclassify real-life adversarial examples [Kurakin *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 result, this approach causes targeted misclassification, which changes a stop sign into a speed limit sign for the AI system. 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] generated 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 "barbershop" with .82% confidence.

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

Figure 1: a) AdvCam and b) RP_2 techniques can deceive CNN classifiers by introducing perturbations. The SL45 refers to the speed limit of 45.

Adaptive Square Attack (ASA) [Li *et al.*, 2020] proposed that it can attack black-box systems by creating imperceptible perturbations for traffic sign images, causing misclassification of the sign. Another study investigated five adversarial attacks and four defence methods on three driving models used in modern AVs [Deng *et al.*, 2020]. The results showed that while the defence methods were effective against several attacks, none could completely protect against all five attacks.

A recent study deceived a DNN-based traffic sign recognition with realistic-looking stickers [Bayzidi *et al.*, 2022]. Another attack included painting the road, which targeted DNN models used in AVs [Boloor *et al.*, 2019]. A successful physical adversarial attack was demonstrated on a commercial classification system to deceive an AV's sign classifier [Morgulis *et al.*, 2019]. BadNets algorithm [Gu *et al.*, 2019] was implemented to deceive a complex traffic sign detection system leading to misclassifying stop signs as speed-limit signs in real-world images.

Concisely, several studies have demonstrated the vulnerability 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 driving systems and highlight the need for more robust and resilient 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.

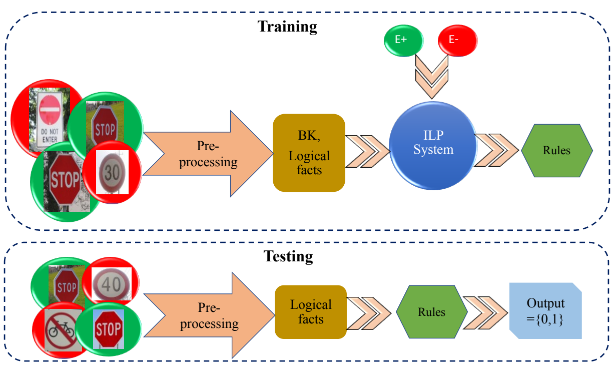

Figure 2: ILP- based traffic sign classifier

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.

Table 1: Aleph Experiment Mode Declarations

$:-modeh(1, traffic\_sign(+sign, \#class)).$
$:-modeb(*, has\_colour(+sign, \#colour)).$
$:-modeb(*, has\_shape(+sign, \#shape)).$
$:-modeb(*, has\_word(+sign, -w)).$
$:-modeb(*, closely\_match(+w, \#word)).$
$:-modeb(*, has\_number(+sign, -n)).$
$:-modeb(*, has\_digits(+n, \#int)).$

Table 2: Metarules utilized in the Metagol-based experiment, predicate symbols (second-order variables) are represented using uppercase letters, while variables are represented using lowercase letters.

Name	Metarule
Identify	P(x,y)⟵ Q(x,y)
Inverse	P(x,y)⟵ Q(y,x)
Precon	P(x, y)⟵ Q(x), R(x, y)
Postcon	P(x, y)⟵ Q(x, y), R(y)
Chain	P(x, y)⟵ Q(x, z), R(z, y)
Recursion	P(x, y)⟵ Q(x, z), P (z, y)

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_w1)* and *closely_match(a_w1, #word)* can appear in the body of the induced rule. The former predicate holds when the sign "a" has the word a_w1 on it, while the second one holds when the word "a_w1" 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". During 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.

Pos example(p1)	Neg example(n1)
$has\_color\,(p1, red)$.	$has\_color\,(n1, red)$.
$has\_color\,(p1, white)$.	$has\_color\,(n1, white)$.
$has\_shape\,(p1, octagon)$.	$has\_shape\,(n1, Circle)$.
$has\_word\,(p1, p1\_w1)$.	$has\_number\,(n1, n1\_d1)$.
$closely\_match\,(p1\_w1, stop)$.	$has\_digits\,(n1\_d1, 30)$.

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 subset 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. a) positive set of stop signs, b) positive set of speed limit 45 signs and c) negative set for both stop and speed limit 45 dataset)

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] is 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, Metagal 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 RP_2 (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 and is presented below:

```
traffic_sign(A, stop_sign):-
    has_word(A, A_w1),
    closely_match(A_w1, stop).
traffic_sign(A, speedsign) :-
    has_number(A,B),
    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-

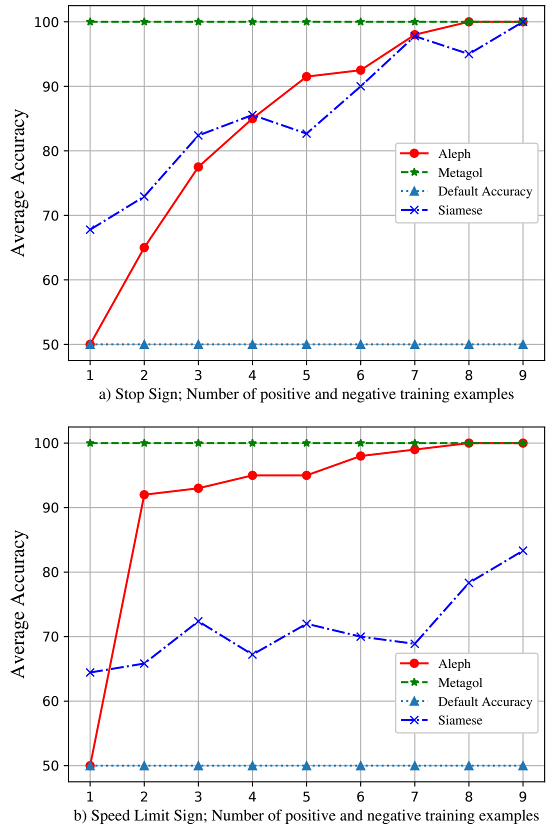

Figure 4: The average accuracy of Aleph, Metagol and the Siamese network versus the number of training examples from the base dataset (equal positive and negative sets) for a) stop sign and b) speed limit sign

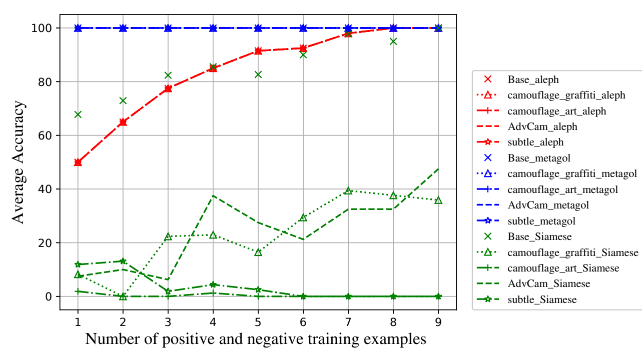

Figure 5: Evaluation of ILP-based systems and Siamese network on different test datasets with different attacks with an increasing number of training examples from the base dataset

Table 4: Summary of comparison of the results of the hypothesis induced by the proposed ILP-based approach (Aleph and Metagol) on different test datasets with the CNN-based classifiers.

dataset		non-compositional		knowledge-
		CNN-based	Siamese	based(ILP)
base		97.6%	100%	100%
RP_2	subtle	0%	0%	100%
RP_2	cam graffiti	0%	35.9%	100%
RP_2	cam art	0%	0%	100%
AdvCam		83.3%	47.5%	100%

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.

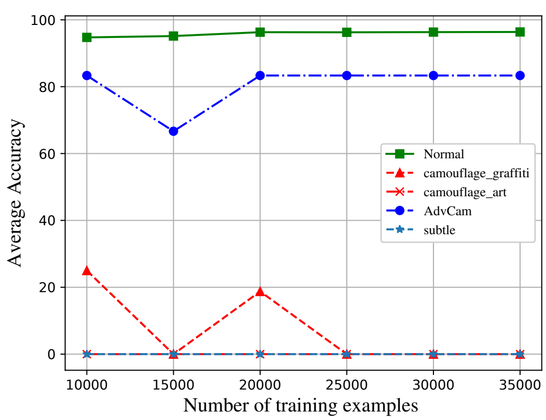

Figure 6: The CNN-based traffic sign classifier evaluation on different datasets

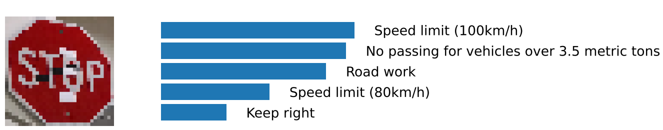

Figure 7: The five top predictions by the CNN-based traffic sign classifier, the traffic sign is targeted with RP_2 (camouflage art) attack.

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.

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

[Anderson *et al.*, 2020] Greg Anderson, Abhinav Verma, Isil Dillig, and Swarat Chaudhuri. Neurosymbolic reinforcement learning with formally verified exploration. *Advances in neural information processing systems*, 33:6172–6183, 2020.

[Arnab *et al.*, 2018] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 888–897, 2018.

[Ashwin Srinivasan, 2001] Ashwin Srinivasan. The aleph manual. https://www.cs.ox.ac.uk/activities/programinduction/Aleph/aleph.html, 2001.

[Bayzidi *et al.*, 2022] Yasin Bayzidi, Alen Smajic, Fabian Hüger, Ruby Moritz, Serin Varghese, Peter Schlicht, and Alois Knoll. Traffic sign classifiers under physical world realistic sticker occlusions: A cross analysis study. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 644–650. IEEE, 2022.

[Boloor *et al.*, 2019] Adith Boloor, Xin He, Christopher Gill, Yevgeniy Vorobeychik, and Xuan Zhang. Simple physical adversarial examples against end-to-end autonomous driving models. In *2019 IEEE International Conference on Embedded Software and Systems (ICESS)*, pages 1–7. IEEE, 2019.

[Chen *et al.*, 2017] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. *arXiv preprint arXiv:1712.02051*, 2017.

[Cropper and Muggleton, 2016] Andrew Cropper and Stephen H. Muggleton. Metagol system. https://github.com/metagol/metagol, 2016.

[Cropper *et al.*, 2020] Andrew Cropper, Sebastijan Dumancic, and Stephen H Muggleton. Turning 30: New ideas in inductive logic programming. In *IJCAI*, 2020.

[Deng *et al.*, 2020] Yao Deng, Xi Zheng, Tianyi Zhang, Chen Chen, Guannan Lou, and Miryung Kim. An analysis of adversarial attacks and defenses on autonomous driving models. In *2020 IEEE international conference on pervasive computing and communications (PerCom)*, pages 1–10. IEEE, 2020.

[Duan *et al.*, 2020] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A Kai Qin, and Yun Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1000–1008, 2020.

[Eykholt *et al.*, 2018a] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Physical adversarial examples for object detectors. *arXiv preprint arXiv:1807.07769*, 1(3):4, 2018.

[Eykholt *et al.*, 2018b] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

[Gu *et al.*, 2019] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[Gui *et al.*, 2021] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering*, 2021.

[Hendrik Metzen *et al.*, 2017] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 2755–2764, 2017.

[Koch *et al.*, 2015] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[Kop, 2020] Mauritz Kop. Machine learning & eu data sharing practices. In *TTLF Newsletter on Transatlantic Antitrust and IPR Developments*. Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust . . . , 2020.

[Kurakin *et al.*, 2018] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[Lee, 2018] Timothy B Lee. Report: Software bug led to death in uber's self-driving crash. *Ars Technica*, 5, 2018.

[Leech *et al.*, 2021] Gavin Leech, Nandi Schoots, and Joar Skalse. Safety properties of inductive logic programming. In *SafeAI@ AAAI*, 2021.

[Li *et al.*, 2020] Yujie Li, Xing Xu, Jinhui Xiao, Siyuan Li, and Heng Tao Shen. Adaptive square attack: Fooling autonomous cars with adversarial traffic signs. *IEEE Internet of Things Journal*, 8(8):6337–6347, 2020.

[Morgulis *et al.*, 2019] Nir Morgulis, Alexander Kreines, Shachar Mendelowitz, and Yuval Weisglass. Fooling a real car with adversarial traffic signs. *arXiv preprint arXiv:1907.00374*, 2019.

[Muggleton *et al.*, 2015] Stephen H Muggleton, Dianhuan Lin, and Alireza Tamaddoni-Nezhad. Meta-interpretive learning of higher-order dyadic datalog: Predicate invention revisited. *Machine Learning*, 100(1):49–73, 2015.

[Muggleton, 1991] Stephen Muggleton. Inductive logic programming. *New generation computing*, 8(4):295–318, 1991.

[Nikolaus *et al.*, 2019] Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. Compositional generalization in image captioning. *arXiv preprint arXiv:1909.04402*, 2019.

[Qin *et al.*, 2020] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020.

[Stallkamp *et al.*, 2012] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Thys *et al.*, 2019] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[Vivek Yadav, 2016] Vivek Yadav. German sign classification using deep learning neural networks. https://github.com/vxy10/p2-TrafficSigns, 2016.

[Yan *et al.*, 2022] Chen Yan, Zhijian Xu, Zhanyuan Yin, Xiaoyu Ji, and Wenyuan Xu. Rolling colors: Adversarial laser exploits against traffic light recognition. *arXiv preprint arXiv:2204.02675*, 2022.