InformatiCup Competition 2019: Fooling Traffic Sign Recognition

Marcus Soll(✉)

Universität Hamburg, Vogt-Koelln-Str. 30, 22527 Hamburg, Germany
2soll@informatik.uni-hamburg.de
https://www.uni-hamburg.de/

Abstract. Neural networks are used more and more in critical areas such as autonomous driving. In such cases, their limitations might cause dangerous situations. Researchers were able to show that such limitations enable attacks on systems containing neural networks, which are even possible in real world scenarios. For example, a state-of-the-art network might misclassify modified traffic signs. Other researchers have shown that modern car assistants can easily be fooled to drive the car into the wrong lane on a street.

The InformatiCup is a collegiate computer science competition in Germany, Switzerland and Austria for all students, with tasks based on real world problems. This year's task is based on the above mentioned problem. To demonstrate this problem and to motivate students for experimenting with neural networks, participants were asked to generate fooling images for a traffic sign classifying neural network without having direct access to the network. The images should not be recognisable by humans as traffic signs, but be classified as such with a high confidence by the neural network.

Keywords: Fooling images · Neural network · Competition

1 Motivation

Imagine driving an autonomous car down a road. You pass a sticker with some weird signs (such a scene might appear as shown in Fig. 1). Suddenly, your car stops in the middle of the road because your car has identified the sticker as a "No through traffic" sign. This way, fooling images lead to dangerous situations.

The troubling thing is that similar attacks have already been shown in practice. For example, Eykholt et al. [4] were able to let neural networks misclassify traffic signs by slightly modifying them. The Tencent Keen Security Lab [14] was able to mislead a Tesla car into the wrong lane by simply placing coloured dots on the road.

To get more insights into fooling images, the participants got the task of creating these images against a target network. The hope was that students can be motivated into experimenting with neural networks through this task.

C. Benzmüller and H. Stuckenschmidt (Eds.): KI 2019, LNAI 11793, pp. 325–332, 2019.
https://doi.org/10.1007/978-3-030-30179-8_29

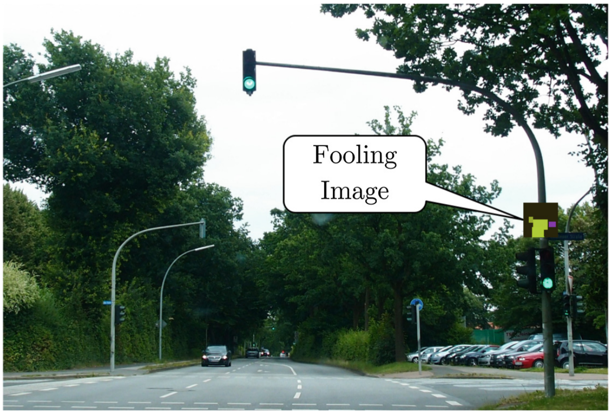


Fig. 1. Example of a fooling image on the road, which might be classified as a "No through traffic" sign.

2 InformatiCup

The InformatiCup[1] (established 2005) is a yearly computer science competition held by the Gesellschaft für Informatik[2] for students in all branches of study in Germany, Switzerland and Austria. It offers prize money to the winning team as well as prizes for the best teams.

The topics are oriented on real world problems. Past topics include, for example, harvesting strategies for manganese nodules or the prediction of fuel prices.

The InformatiCup is a competition with a holistic approach, where the whole solution is important and not only the programming or simply the quality of the results. The judgement is based on the following criteria:

- the theoretical background of the solution
- quality of programming (including software architecture and quality management)
- presentation
- quality of the result (e.g. accuracy)
- user manual
- additions by the teams (like graphical user interfaces)

The InformatiCup is well recognised in the industry and some tasks have even contributed to research [11,12].

3 Background: Fooling Images

In recent research it was shown that neural networks, despite showing high accuracy for many tasks including images classification [7], are susceptible to malicious input. Most research focuses on so called *adversarial examples* [1]. In these

[1] https://gi.de/informaticup/

[2] https://gi.de/

examples, noise is added to a correctly classified image to make neural networks misclassify it, while at the same time the change should not be detectable by humans.

Another approach was taken by Nguyen et al. [8] and followed by Soll [10]. In contrast to adversarial examples, *fooling images* are not created from existing images. Instead, they are created artificially and are classified by neural networks with high confidence, while not being recognisable by humans.

4 Task Description

The task of the InformatiCup 2019 was to develop a software solution that is able to generate fooling images for a provided neural network for at least five different traffic signs. All generated images must be classified as a traffic sign by the neural network with at least 90% confidence. There were no requirements of a specific traffic sign, which allowed untargeted attacks. Some examples of possible solutions (generated by the jury) can be seen in Fig. 2.

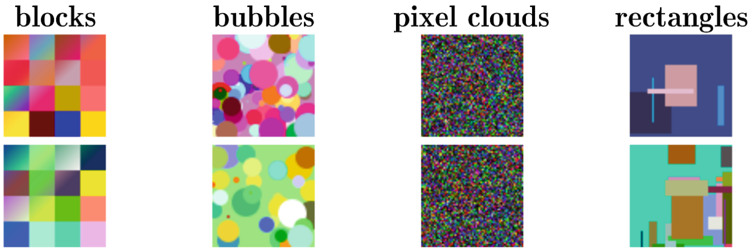


Fig. 2. Examples of fooling images created by the organisers of the competition. All are classified as a "Vorfahrt" sign (priority traffic sign) with at least 90% confidence by the target network. All images are 64 × 64 pixels in size.

To make the task harder, several limitations were put in place for the participants:

- The participants had no direct access to the network. Instead, a Web API had to be used.
- The neural network architecture was unknown to the participants.
- Each team was restricted to only 60 requests per minute. This was checked by providing each team a unique API key.
- The number of output classes was not provided to the participants, although the dataset was known.
- Only the top 5 classification results were returned.

In addition to the software solution, teams had to turn in a paper describing the theoretical background, the software design decisions and a result discussion.

A user manual was required, which could be part of the paper or a separate paper. The best teams had to present their solution to a jury consisting of members from industry and academia.

4.1 Neural Network Used

For the task, a simple single-layer neural network (see Fig. 3) was trained on *The German Traffic Sign Recognition Benchmark* [13]. The input of the network was an image with 32 × 32 pixels. In an attempt to reduce the susceptibility of the neural network against fooling images, several measures were taken:

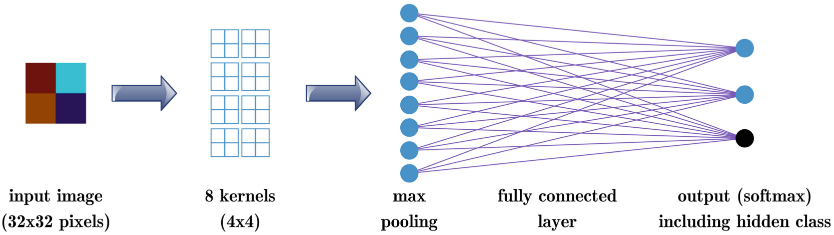


Fig. 3. Schema of neural network architecture.

1. The architecture was deliberately simple, with only one layer and eight kernels. This was to keep the Vapnik-Chervonenkis-dimension and with it the required data for optimal learning low (see [6]).
2. In addition to the provided images in the dataset, additional images were generated and used as a hidden class (i.e. not visible to the participants) in the training (see Fig. 4). The goal was to make the network more robust to certain changes. These images include:
   - Images with a single colour to counter background detection
   - Images with random circles to counter the shape of the traffic signs
   - Images with random noise to counter reaction to noise
3. The network was trained on a low number of epochs (five) to counter overfitting.

With all those measurements, the network reached an accuracy of about 85%. Although the accuracy might not be as high as desired, it seemed suitable enough for the competition.

To ensure that the network is not easily fooled, a dataset of 38 images (including 10 random noise images) was tested against the neural network, of which none was detected as a traffic sign.

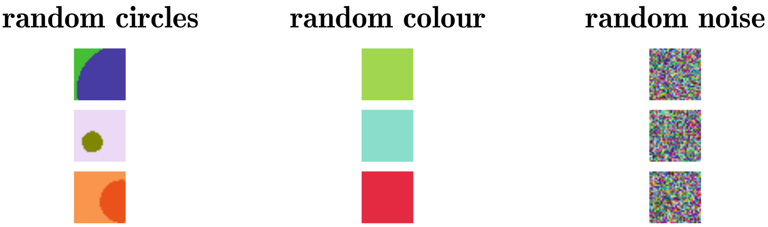
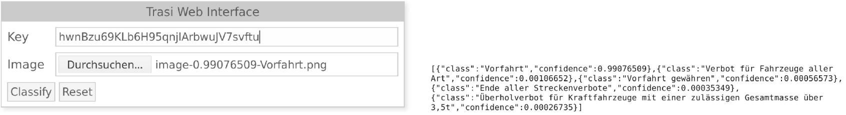


Fig. 4. Examples of images added as a hidden class to the trainings set to reduce the susceptibility against fooling images. (Color figure online)

4.2 Network Interface

For the communication with the neural network, two interfaces were provided.

Website: Through a website, a single image could be classified. A view on the website is shown in Fig. 5.



Key hwnBzu69KLb6H95qnjlArbwuJV7svftu

Image image-0.99076509-Vorfahrt.png

[{"class":"Vorfahrt","confidence":0.99076509},{"class":"Verbot für Fahrzeuge aller Art","confidence":0.00106652},{"class":"Vorfahrt gewähren","confidence":0.00056573},
{"class":"Ende aller Streckenverbote","confidence":0.00035349},
{"class":"Überholverbot für Kraftfahrzeuge mit einer zulässigen Gesamtmasse über 3,5t","confidence":0.00026735}]

(a) A single image can be uploaded with a valid API key.

(b) Example feedback of the website.

Fig. 5. Website as an interface to the neural network.

Web API: A HTTP POST-Request in the encoding *multipart/form-data* containing the API key and the image could be sent to an endpoint. If the request is valid, a JSON object containing the top five prediction and the corresponding confidence values is returned.

5 Analysis of the Competition

In this year's competition, 46 teams from all over Germany registered for the competition. Out of those, 30 teams turned in a working final solution, which were all graded by a jury from both industry and academia. Out of those, five teams advanced to the final round and presented their solutions. The results (including links to repositories of the student solutions) can be found on GitHub[3].

This year's competition not only had the highest number in solutions turned in (for comparison, the InformatiCup 2018 had a total of 17 teams turning in

[3] https://github.com/InformatiCup/InformatiCup2019/blob/master/results/README.md

solutions), but also the solutions were of high quality. This shows that the interest in artificial intelligence/neural networks and their limitations was high. This was also confirmed by the participants in personal discussions at the final round.

The participating teams turned in a wide variety of different solutions. The winners combined methods from recent research to generate fooling images: They trained a new substitution model based on the feedback from the Web API plus the provided dataset (based on Papernot et al. [9] with slight modifications) and applied state-of-the-art methods from adversarial examples research on the new model (Modified attack of Carlini and Wagner [2] as well as Eykholt et al. [4]).

The team in second place provided two solutions:

- Based on an initial image (e.g. black background), the image is divided into blocks with a user defined size. For each block (in random order), the effect of the different colours is analysed. The colour variant with the highest confidence is chosen for that block. The algorithm ends when the target confidence is reached.
- Similar to the winning team, the team trained a substitution model (based on Papernot et al. [9]). For this, they used 1000 randomly chosen images of the dataset for training (100 for the test set), and used the Jacobian-based Data Augmentation [9] for training. They then used the Momentum-based iterative Fast Gradient Sign Method [3] for generating fooling images (either targeted for a specific sign or untargeted).

The team in third place used genetic algorithms [5,15] for creating untargeted fooling images, however they omitted the recombination phase of the genetic algorithm and only used mutation. Starting on a user provided image, they implemented different mutation methods (all controlled by variables modifiable by the user):

- Set a percent of pixels to a random colour.
- Draw circles (either filled or unfilled) on the image.
- Draw rectangles (either filled or unfilled) on the image.
- Draw multiple polygons (always filled) on the image in a way that results in rotationally symmetrical placement of the polygons.
- Divide the image into a grid of blocks and slightly stain the different blocks with random colours (keeping the original image visible if desired).

Besides the three approaches described here, many more solutions were turned in with vastly different approaches. However, it is not in the scope of this paper to describe all approaches in detail.

6 Conclusion

Neural networks are widely used. However, their limitations - like fooling images - are not understood well. The InformatiCup used this as a topic of this year's competition, where participants should generate fooling images for a traffic sign classifying neural network. With 30 solutions from all over Germany, this year's competition has motivated students to look into neural networks. The InformatiCup will be continued in 2020.

Acknowledgments. I would like to thank all organisers, jury and participants of the InformatiCup 2019 competition organised by the “Gesellschaft für Informatik”. A special thanks goes to the sponsors of the event: Amazon, PPI AG, Netlight, Volkswagen, TWT GmbH Science & Innovation and GitHub. Further thanks goes to the AutoUni Wolfsburg for hosting the final presentation.

References

1. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access **6**, 14410–14430 (2018). https://doi.org/10.1109/ACCESS.2018.2807385
2. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57, May 2017. https://doi.org/10.1109/SP.2017.49
3. Dong, Y., et al.: Boosting adversarial attacks with momentum. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 9185–9193 (2018). https://doi.org/10.1109/CVPR.2018.00957
4. Eykholt, K., et al.: Robust physical-world attacks on deep learning visual classification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1625–1634, June 2018. https://doi.org/10.1109/CVPR.2018.00175
5. Gerdes, I., Klawonn, F., Kruse, R.: Evolutionäre Algorithmen: Genetische Algorithmen – Strategien und Optimierungsverfahren – Beispielanwendungen. Vieweg+Teubner Verlag, Wiesbaden (2004). https://doi.org/10.1007/978-3-322-86839-8
6. Harman, G., Kulkarni, S.: Statistical learning theory as a framework for the philosophy of induction. In: Bandyopadhyay, P.S., Forster, M.R. (eds.) Philosophy of Statistics. Handbook of the Philosophy of Science, vol. 7, pp. 833–847. North-Holland, Amsterdam (2011). https://doi.org/10.1016/B978-0-444-51862-0.50027-7. http://www.sciencedirect.com/science/article/pii/B9780444518620500277
7. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deepneural network architectures and their applications. Neurocomputing **234**, 11–26 (2017). https://doi.org/10.1016/j.neucom.2016.12.038. http://www.sciencedirect.com/science/article/pii/S0925231216315533
8. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 427–436, June 2015. https://doi.org/10.1109/CVPR.2015.7298640
9. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, ASIA CCS 2017, pp. 506–519. ACM, New York (2017). https://doi.org/10.1145/3052973.3053009
10. Soll, M.: Fooling deep neural networks using Cuckoo Search. Technical report, University of Hamburg, February 2016. https://doi.org/10.13140/RG.2.1.1402.7760
11. Soll, M., Naumann, P., Schöning, J., Samsonov, P., Hecht, B.: Helping computers understand geographically-bound activity restrictions. In: Proceedings of the 2016

CHI Conference on Human Factors in Computing Systems, CHI 2016, pp. 2442–2446. ACM, New York (2016). https://doi.org/10.1145/2858036.2858053

12. Soll, M., Vosgerau, M.: ClassifyHub: an algorithm to classify GitHub repositories. In: Kern-Isberner, G., Fürnkranz, J., Thimm, M. (eds.) KI 2017. LNCS (LNAI), vol. 10505, pp. 373–379. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67190-1_34
13. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural Netw. **32**, 323–332 (2012). https://doi.org/10.1016/j.neunet.2012.02.016. http://www.sciencedirect.com/science/article/pii/S0893608012000457. Selected Papers from IJCNN 2011
14. Tencent Keen Security Lab: Experimental security research of tesla autopilot. Technical report, Tencent Keen Security Lab (2019)
15. Weicker, K.: Evolutionäre Algorithmen. Springer Fachmedien Wiesbaden, Wiesbaden (2015). https://doi.org/10.1007/978-3-658-09958-9