Automated Method for Evaluating Neural Network's Attention Focus

Tomasz Szandała[1][0000-0003-4525-0444] and Henryk Maciejewski[1][0000-0002-8405-9987]

[1] Wroclaw University of Science and Technology, Wroclaw, Poland
Tomasz.Szandala@pwr.edu.pl
Henryk.Maciejewski@pwr.edu.pl

Abstract. Rapid progress in machine learning and artificial intelligence (AI) has brought increased attention to the potential security and reliability of AI technologies. This paper identifies the threat of network incorrectly relying on counterfactual features that can stay undetectable during validation but cause serious issues in life application. Furthermore, we propose a method to counter this hazard. It combines well-known techniques: object detection tool and saliency map obtaining formula to compute metric indicating potentially faulty learning. We prove the effectiveness of the method, as well as discuss its shortcomings.

Keywords: deep learning, xai, convolutional neural networks, region proposal, object detection.

1 Introduction

Object recognition networks are trained to minimize the loss on a given training dataset and subsequently evaluated with regard to the testing dataset. With this approach, we expect the model will learn its own correlation between images and their labels. However, the actual model reliability and robustness highly depend on the choice of the unskewed and heterogenous training dataset. Otherwise, we may encounter a latent correlation that is present in our training set but is not a reliable indicator in the real world.

Such unsought correlation can be color yellow for a cab or biased background representing a typical environment for classified objects like a desert for the camel or sea for the ship. Multiple works [10, 11] prove that background bias can have a serious impact on classifier reliability and even state-of-the-art models are not free of relying on latent correlations [16]. Yet these defects are recognizable merely by manual examination. While for several classes it is achievable, for a wider range of types it can be challenging.

In this paper we aim to provide a solution to detect and avert learning misguided correlations. For this purpose, we choose a pre-trained tool, Detectron2 [2,6], to indicate potential Regions of Interest (ROI) on a given image. Secondly, we process images through our to-inspect network and thus obtain saliency maps with the

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33

GradCAM technique[3] for the most accurate class. Since the GradCAM result is a matrix that assigns potential importance for a given classification to areas of the image, we can calculate the importance value inside the detected ROI and divide it by the entire image's importance. The obtained ratio is considered a fitting metric for a given classification. The higher it is, the more certain we can be that the network has learned the actual object, not the background.

During our research, we have performed 4 experiments. In the first one, we have provided the evidence for the problem. Secondly, we have formulated the method details and tested it on an ImageNet's subset of 9 classes. This proved the effectiveness of the proposed method and, due to the small set, we could easily verify its effectiveness. In the third attempt, we moved to the entire ImageNet set and pretrained VGG-11 from Model Zoo [4]. In this phase, we used the method to determine several classes with potential faulty correlations. We have further investigated them and in the final experiment we used the possessed knowledge to create a successful adversarial attack on a presumably robust network [5].

2 State of the art

Image backgrounds are a natural source of correlation between images and their labels in object recognition. Indeed, prior work has shown that models may use backgrounds in classification [14, 16, 17, 18, 19], and suggests that even human vision makes use of image context for scene and object recognition [20].
Moreover, Xiao et al. [16] have demonstrated that standard models not only use but require backgrounds for correctly classifying large portions of test sets. They have also studied the impact of backgrounds on classification for a variety of classifiers, and found that more accurate models tend to simultaneously exploit background correlations more and have greater robustness to changes in the image background.
The researchers from Zhang's team [7] created a CNN that classifies an X-ray image of the chest as "high disease probability" based not on the actual appearance of the disease but rather on an "L" metal token placed on the patient's left shoulder. The key is that this "L" token is placed directly on the patient's body only if he is lying down, and the patient will only lie down on the X-rays if they are too weak to stand. Therefore, CNN has learned the correlation between "metal L on the shoulder" and "patient is sick" - but we expect CNN to be looking for real visual signs of the disease, not metal tokens.
The final example: CNN learns [13] to classify an image as a "horse" based on the presence of a source tag in the lower-left corner of one-fifth of the horse's images in the data set. If this "horse source tag" is placed on a car image, the network classifies the image as "horse".

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33

3 Method

The proposed method is based on three simple concepts. The first is to determine what network should consider as a discriminative feature. Secondly, we have to go through and discover what the network is actually learning. And finally, we need to evaluate how much of the learned correlations are tied to the expected area of the image.

3.1 ROI generation

The first phase is the generation of the Regions-of-Interest (ROI). For this purpose, we can employ an expert that will manually indicate ROI, but in this paper we utilized the Detectron2 framework [6]. It is a Facebook AIResearch's next-generation software system that implements state-of-the-art object detection algorithms. It is also a common practice to use a base model pre-trained on a large image set (such as ImageNet [9]) as the feature extractor part of the network. Detectron2 framework allows us to obtain coordinates of two points that mark the top left and bottom right of the ROI. For the purpose of this research, we did not put attention to which class object has been recognized.

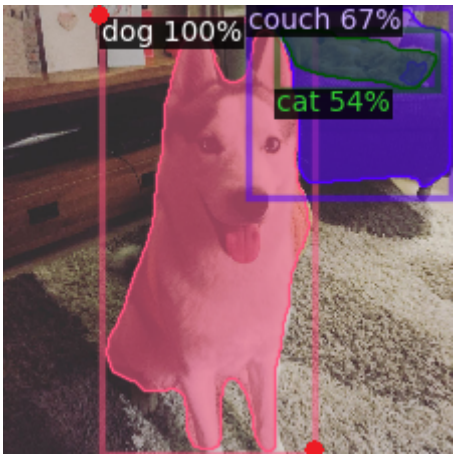


Fig. 1. Detectron2 processed image with highlighted 3 regions of interests

Detectron2 produces images with highlighted ROIs (see fig. 1) and returns a tuple for each image that contains 4 coordinates: x and y for the top left corner and x and y for the bottom right corner of the ROI.

3.2 Saliency map generation

Saliency map generation is the subsequent stage, although it can be performed in parallel to ROIs generation if we have sufficient computing power. For this purpose

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33

we have used the popular GradCAM method[3]. To obtain the class-discriminative localization map, Grad-CAM computes the gradient of y^c (score for class c) with respect to feature maps A of a convolutional layer. These gradients flowing back are global-average-pooled to obtain the importance weights for pixel k with respect to class c: α^c_k (eq. 1).

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}} \qquad (1)$$

The importance weights create a saliency map as a matrix of values between 0.0 to 1.0 which corresponds to the importance of a particular pixel. Graphical representation of this map is a picture with colors from blue (lowest importance) to red (highest importance).

3.3 Fit metric calculation

The final step is to compute how many classification-important pixels are inside ROI in relation to all important pixels in the image. In the proposed method we sum values of saliency map inside ROI and divide them by the sum of values over the entire map. If there are more ROIs detected we calculate the ratio for each separately and in the end choose the highest value.

$$m_{fit} = \frac{\sum_{x,y}^{roi} \alpha_k^c}{\sum_{x,y}^{image} \alpha_k^c} \qquad (2)$$

This proportion gives us fit measurement. The higher value we obtain for a given class, the higher confidence that the network has learned the object, not the background.

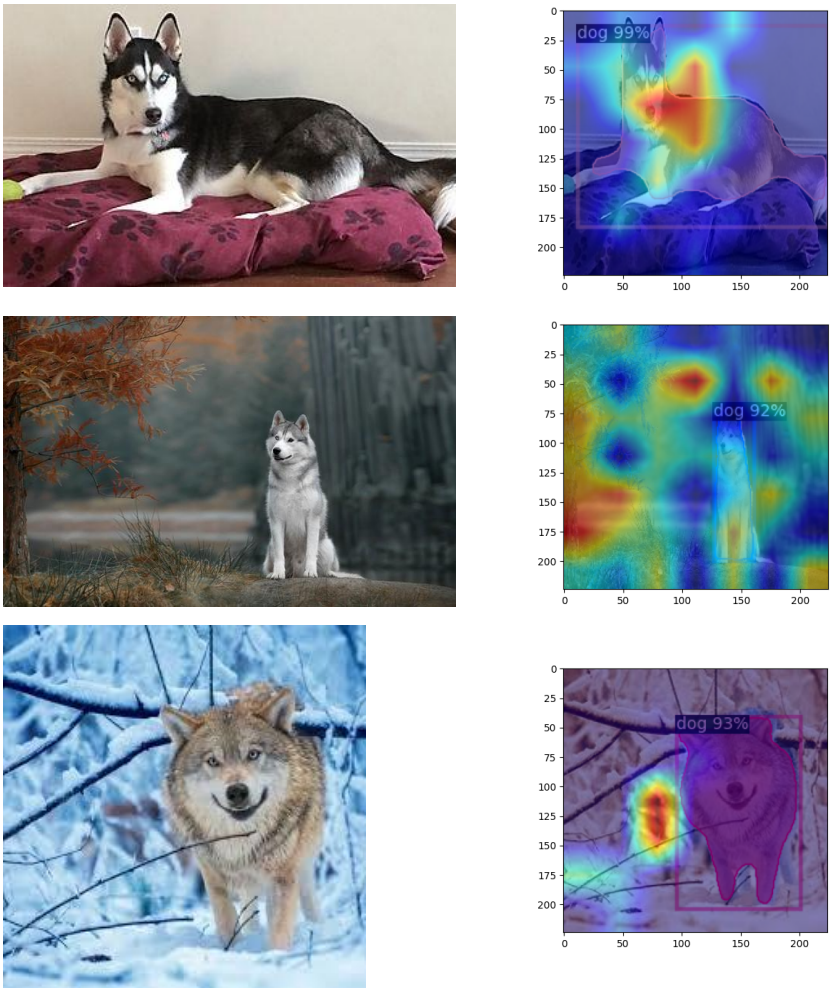
4 Experiments

Our research has been divided into 4 linked experiments.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33

4.1 Huskies and Wolves discrimination using ResNet architecture

In the first study, we are trying to reproduce the Singh's et al. [13] experiment about discrimination of wolves and huskies. They have noticed that their network has learned snow as a feature strictly correlated with wolves' class.
In our attempt, we have collected 40 images of each specimen and taught the network as a binary classifier. This time attention has been paid to ensure that wolves are not only displayed in the winter background. The precision achieved was satisfactory. However, this time the visualization methods showed that most of the wolf images were taken in their natural habitat, which usually contained bushes, trees and branches, and these were the criteria for distinguishing the wolf from the husky.



ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33

Fig. 2. Top-down: husky classified as husky and its saliency map, husky classified as wolf due to trees and branches, wolf classified as wolf due to branches

The features that determine wolves according to the network are leafless trees and bushes (fig. 2.). Indeed, most of our wolves are pictured in their natural habitat and often with some trees in the background. On the other hand, the huskies are all beasts without any branches.
This research supports the thesis that networks may learn incorrect correlations even when the training set appears to be hardened against this setback. Nonetheless, when there are only 2 classes the flaw can be easily exposed using known visualization techniques.

4.2 Re-trained ResNet50 on ImageNet's subset with 9 classes

For this research, we have utilized transfer learning [12] of ResNet50 and expect it to recognize 9 subjectively chosen classes. The subset ImageNet-9 consists of classes: bird, camel, dog, fish, insect, musical_instrument, ship, snowboard, wheeled_vehicle. Each class is often associated with certain environments like camel-desert, insect-plants, ship-water, snowboard-mountains, etc. Each class has been represented by approximately 500 images. It appears that transfer learning is quite an efficient method since 8 object types were recognized correctly as the actual object. Only one class, snowboard, took our attention. It appears that the presence of the sky implicated this recognition.
Using CAM-in-ROI method we have noticed that only 28% of significant pixels were inside ROI's rectangle, while for e.g. camels this value oscillated around 74%.

Table 1. ROI-fitting value and percentage of correctly classified images over ImageNet-9.. Note that images where ROI has been not found were excluded

Class	Average fitting	Model accuracy
bird	77.75 %	88.25 %
camel	74.20 %	22.70 %
dog	85.28 %	96.18 %
fish	67.87 %	62.82 %
insect	66.31 %	77.10 %
musical_instrument	59.26 %	81.80 %
ship	68.19 %	88.08 %
snowboard	28.01 %	82.28 %
wheeled_vehicle	83.77 %	94.26 %

Table 1 shows average importance, according to GradCAM, found inside ROI. The most outstanding is the snowboard class where less than one-third of saliency fits

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33

inside indicated ROI. Human conducted scrutiny revealed that for this certain class a top area associated with the sky is signal classification as the snowboard (fig. 3).

It is worth noticing that despite a low fitting ratio for the snowboard, the model's classification accuracy does not differ significantly from other classes. This leads to a conclusion that incorrect correlation may be undetectable by standard means.

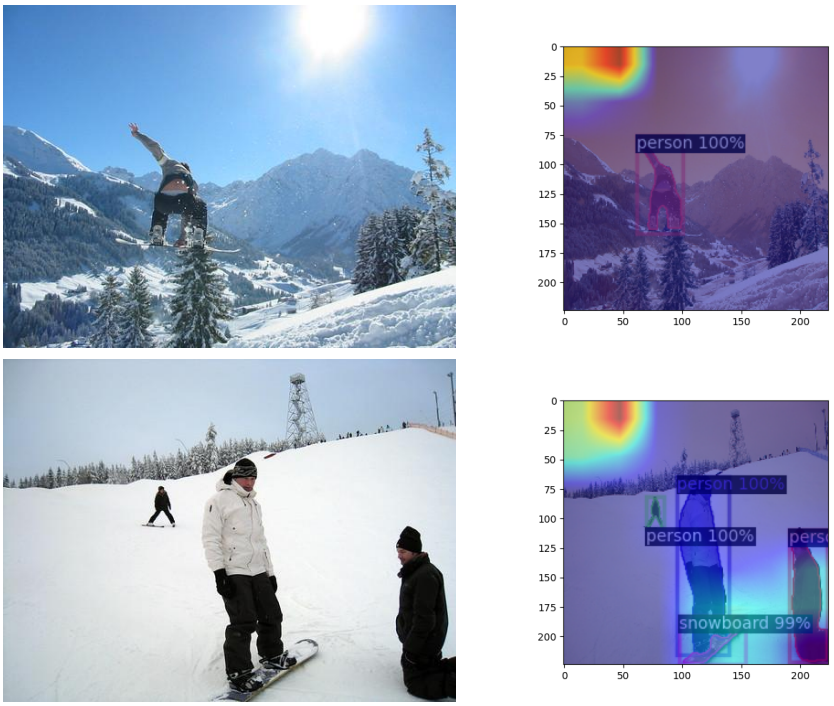


Fig. 3. Images classified as snowboard and their respective saliency maps with ROIs

This proof of concept demonstrated the practicality of the proposed method on a small subset, where a human validation could be done. The method has correctly identified the snowboard classification as defective and other correlation as satisfactory.

4.3 Pretrained VGG-11 for object classification

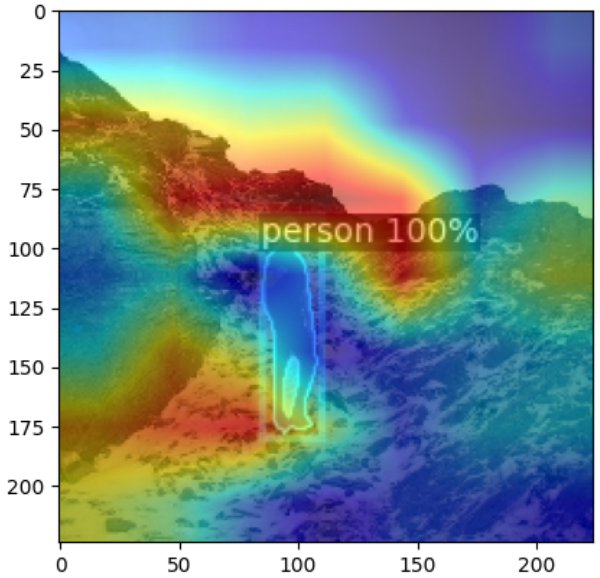
Xiao et al. stated that state-of-the-art pretrained networks appear to be immune to background skewed learning [16]. We have decided to scrutinize this statement by applying CAM-in-ROI measurement to the samples of all classes found in the ImageNet dataset and their classification using VGG architecture. We have prepared a batch of approximately 50 images of each class listed in the ImageNet set. Chosen results are displayed in table 2.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33

Table 2. ROI-fitting value over full ImageNet. Only selected classes were displayed

Class	Average fitting
Japanese_spaniel	96.73%
Persian_cat	93.04%
black_and_gold_garden_spider	11.08%
valley	7.75%
lakeside	7.50%
seashore	2.46%
bell_cote	1.53%
alp	1.12%
breakwater	0.68%
mosque	0.25%
rapeseed	0.00%

Some of the classes, like seashore, alp or lakeside can be considered object-less, since Detectron2 hardly ever recognized any object on the picture. Or there could have been highlighted an object that has only subsidiary importance. As we may see on the fig. 4: ROI targets a person, while the network focuses, correctly, on the surroundings and classifies the image as a valley. Furthermore, some objects like bell cote are not known to the Detectron2. These issues caused many false positives in our results, although these may be overcome by applying different, more compelling ROI detecting tools.



ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33

Fig. 4. An image classified correctly as a valley, despite the person detected in the foreground. Note that a tiny saliency region is included inside ROI thus giving as non-zero fitting measure

Further we must acknowledge one more setback: our technique will turn out ineffective for detecting incorrect perception of the object. Namely like the aforementioned association of the cab object with yellow [21].
Nonetheless, our method has greatly limited the number of classes that require a closer look. Therefore we had to manually inspect only the lowest fitting classes. By reviewing images from the lowest fits we can highlight one distinctive class: black_and_gold_garden_spider. The black-and-gold-garden-spider often appears on the spider web background. While ROI indicates the insect quite well, the saliency map marks the spider web as the most significant area on the image, therefore resulting in only 11% average fitting. This leads us to the conclusion that the network assigns the label black_and_gold_garden_spider to an image that displays an object on a spider's web background.

4.4 **Mocking spider image**

Enhanced with knowledge about background influence on a certain class we can forge an image that has a spider's web in the background and something else in the foreground. We have downloaded a web image as well as a deer's head and merged them into one picture. Before the fusion, each picture was separately correctly classified: web as web and deer as a hartebeest (fig. 5). The blended image, as expected, has been classified as black-and-gold-garden-spider.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33

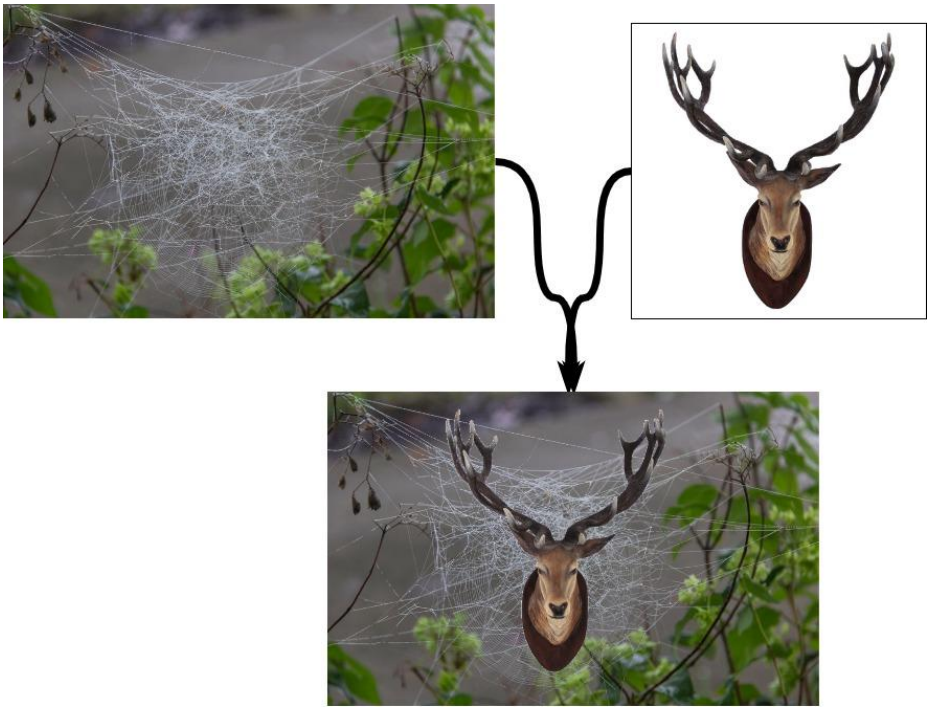


Fig. 5. Source images of spider's web and a deer's head that, according to results from 3rd experiment, should result in classification as black_and_gold_garden_spider

By knowing the flaws in the network's reasoning, we were able to conduct an adversarial attack on it. Its success ultimately proves the potency of the proposed method.

5 Conclusion

Presented evidence proves the existence of the serious problem in deep learning when a network can accidentally learn an incorrect correlation for a given class. Pretrained VGG-11 network from PyTorch's Model Zoo is flawed for at least one class: the black-and-gold-garden-spider.
The proposed method allows neural network practitioners to recognize incorrect correlations picked by the model and thus counter them by e.g. enhancing our training set.
Despite the named drawbacks, the proposed method might be a milestone in improving convolutional neural networks. It allows user to indicate classes that may lack legitimate feature selection and therefore cause complication in real-life applications.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33

6 Future works

The proposed method is currently a proof of concept. It relies on the GradCAM saliency map and detectron2 framework trained on the COCO dataset. Possible improvements consist of choosing a better ROI detection method that knows more types of objects. Moreover, a more complex ROI area could be used instead of a simple rectangular space.
Consequently, the saliency maps generating method can be changed. Possible alternatives are GradCAM++ [3], occlusion maps[8] or excitation backpropagation [7].
Last but not least a refining the formulation of fitting metric can be changed. Currently, it is a naive ratio of in-ROI to entire pixels saliency values. A more accurate measure might rely on integrals computation.

References

1. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Communications of the ACM 60.6 (2017): 84-90.
2. Pham, Vung, Chau Pham, and Tommy Dang. "Road Damage Detection and Classification with Detectron2 and Faster R-CNN." arXiv preprint arXiv:2010.15021 (2020).
3. Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.
4. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
5. Dinh, Vu, and Lam Si Tung Ho. "Consistent feature selection for analytic deep neural networks." arXiv preprint arXiv:2010.08097 (2020).
6. Detectron2, “Detectron2 modelzoo” https://github.com/facebookresearch/detectron2/, accessed:2020-10-16.
7. Zhang, Jianming, et al. "Top-down neural attention by excitation backprop." International Journal of Computer Vision 126.10 (2018): 1084-1102.
8. Shokoufandeh, Ali, Ivan Marsic, and Sven J. Dickinson. "View-based object recognition using saliency maps." Image and Vision Computing 17.5-6 (1999): 445-460.
9. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet:A large-scale hierarchical image database,” in2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255
10. Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
11. Engstrom, Logan, et al. "Adversarial robustness as a prior for learned representations." arXiv preprint arXiv:1906.00945 (2019).
12. Schott, Lukas, et al. "Towards the first adversarially robust neural network model on MNIST." arXiv preprint arXiv:1805.09190 (2018).
13. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.
14. Lapuschkin, Sebastian, et al. "Unmasking Clever Hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1096

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33

15. Zech, John R., et al. "Confounding variables can degrade generalization performance of radiological deep learning models." arXiv preprint arXiv:1807.00431 (2018).
16. Xiao, Kai, et al. "Noise or signal: The role of image backgrounds in object recognition." arXiv preprint arXiv:2006.09994 (2020).
17. Zhang, Jianguo, et al. "Local features and kernels for classification of texture and object categories: A comprehensive study." International journal of computer vision 73.2 (2007): 213-238.
18. Zhu, Zhuotun, Lingxi Xie, and Alan L. Yuille. "Object Recognition with and without Objects." arXiv preprint arXiv:1611.06596 (2016).
19. Sagawa, Shiori, et al. "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization." arXiv preprint arXiv:1911.08731 (2019).
20. Torralba, Antonio. "Contextual priming for object detection." International journal of computer vision 53.2 (2003): 169-191.
21. Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." arXiv preprint arXiv:1811.12231 (2018).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77964-1_33