Adversarial Examples with Specular Highlights

Vanshika Vats[1, 2] and Koteswar Rao Jerripothula[2]

[1]University of California, Santa Cruz
[2]Indraprastha Institute of Information Technology, Delhi
vvats@ucsc.edu, koteswar@iiitd.ac.in

Abstract

We introduce specular highlight as a natural adversary and examine how deep neural network classifiers can get affected by them, resulting in a reduction in their prediction performance. We also curate two separate datasets, ImageNet-AH with artificially generated Gaussian specular highlights and ImageNet-PT by flashing natural specular highlights on printed images, both demonstrating significant degradations in the performance of the classifiers. We note around 20% drop in the model prediction accuracy with artificial specular highlights and around 35% accuracy drop in torch-highlighted printed images. These drops indeed question the robustness and reliability of modern-day image classifiers. We also find that finetuning these classifiers with specular images does not improve the prediction performance enough. To understand the reason, we finally do an activation mapping analysis and examine the network attention areas in images with and without highlights. We find that specular highlights shift the attention of models which makes fine-tuning ineffective, eventually broadly leading to performance drops.

1. Introduction

In recent years, deep learning has allowed us to make significant improvements in image recognition and classification tasks. However, deep neural networks can fail unpredictably. Adversarial attacks make it worse. An adversarially affected image hardly appears any different from the one getting predicted correctly, but it still gets misclassified. Recent studies have shown that adding carefully constructed noise to an image can make it an adversarial image that can fool the network with high confidence [31, 15]. Another interesting work demonstrated that the printed versions of some images could also act as adversarial examples when they ran it through an Inception v3 classifier [24]. A neural network can also be fooled by simply adding an adversarial patch [5] to the image regardless of its scale or location. In [11], Robust Physical Perturbations (RP2) were introduced to prove how images can get affected under physical-world adversary. These studies clearly demonstrated that deep neural networks are far from perfect, and it does not take much effort to deceive a network.

Since ImageNet [9] dataset is considered a benchmark for the image classification [23] and object detection tasks [21], many researchers have been continually working on developing improved models every year to improve the classification performance [17] on this huge dataset, which is supposed to cover all kinds of images. However, more recent studies have shown that even such models are not devoid of getting affected by the adversarial attacks. Models trained on ImageNet can be fooled by adding a small noise vector imperceptible to the human eye using fast gradient sign method [15], by exploiting image semantics to selectively modify colors [28], and by simply adding a perturbation vector without the need of special optimization or gradient computation [25]. Different from these intentional mathematical perturbations, there are also some natural adversaries that can happen. For example, shadows [34] and specular highlights are quite common ones. In this paper, we introduce and analyse specular highlights, which are relatively understudied as a natural adversary and examine how modern-day classifiers can get fooled in their presence.

A specular highlight is a bright spot of light that is observed when a surface reflects off light from a source in a mirror-like fashion. Specular reflection occurs when the angle of reflection of light on a surface is equal to the angle of incidence, i.e., the surface normal is bisecting the angle between the incoming light and the viewer's direction. Thus, the highlighted spot on the object's surface can be directly perceived as somewhat of an image of the light source it is reflecting. Computation of tasks such as image segmentation [1], clustering, recoloring, and object detection [3] can get fooled by the interference of specular highlights. Most modern-day algorithms consider specular highlight regions

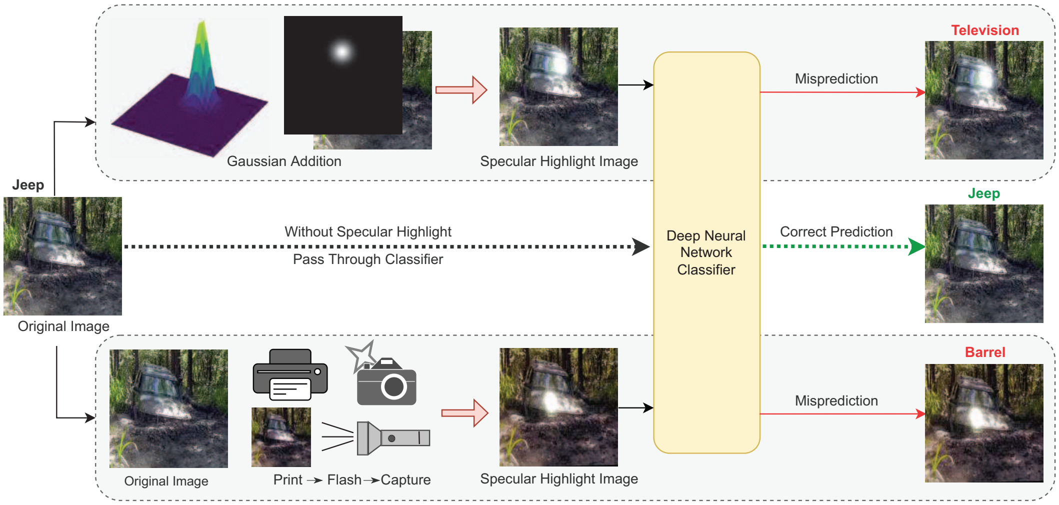

Figure 1. An example of how specular highlights, whether simulated mathematically as Gaussian speckle or thrown naturally by a torch, can affect the classification performance of a deep neural network causing it to make wrong predictions.

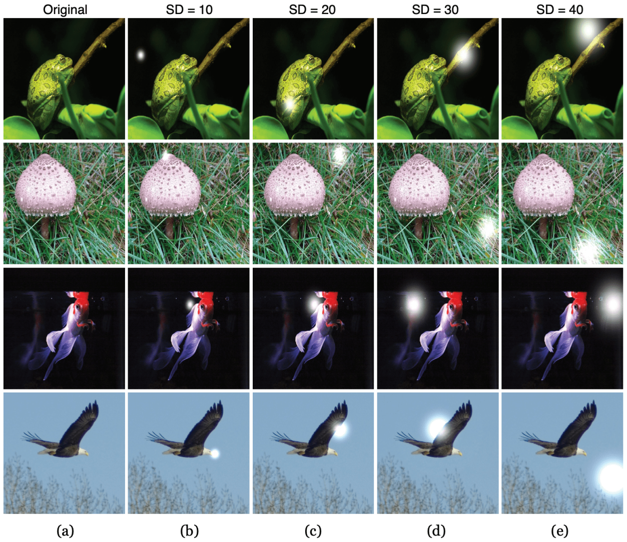

Figure 2. Augmented specular highlight generation to build ImageNet-AH. The figure shows sample images affected by mathematically generated Gaussian specular highlight at varying locations with varying standard deviations (SD).

on an object or image as outliers and only consider perfect diffuse surfaces to make their decisions [2]. This assumption of simplification affects the robustness and applicability of algorithms negatively. Some vision algorithms rely heavily on saliency [33], [12]. Specular highlights can be very salient at times resulting in shift of saliency from the main object to themselves, or on other regions, and hence leading to incorrect results. In real-world scenarios, too, the specular highlights can limit the computation power of the vision algorithms. For instance, glare on the road signs at night can confuse a computer vision sign detection algorithm in autonomous vehicles. Perception algorithms can be fooled by glare on the objects even at sunset/sunrise [10]. As can be seen in Fig. 1, a mathematically generated glare on the windshield and a naturally thrown specular highlight on the bonnet of a jeep are making the image being misclassified as a 'Television' and 'Barrel' respectively. We will also observe in this study that the modern-day classifiers fail even though the highlight is in the background and not directly on the subject. Therefore, it becomes necessary to consider a specular highlight as a natural adversary and examine its effect on the vision tasks.

Thus, we introduce a study of how often a specular highlight can affect a deep neural network (DNN) classifier's performance, bringing out the need to revisit our understanding of the faults of a typical classifier. We do this by not considering this advesary as an 'attack', but an often unnoticed natural phenomenon affecting the performance of algorithms. We first demonstrate that the specular highlights act adversarially on various known DNN classifiers by introducing multiple artificially generated specular highlights on images with varying locations and sizes. We then confirm it through collection of images which have specular highlight generated by flashing of torch on printed images. We further analyze if finetuning the networks with the previously failed specular images will be helpful to improve the performance of the classification. Finally, we do an activation mapping analysis on both the original and specular highlight affected images and find the attention shifts somewhat responsible for the model mispredictions. Earlier, Hendrycks et al. [18] did introduce natural adverasary as something to bother about, but they don't specifically identify particular adversaries and try to address them. Therefore, through this work, we wish to begin a novel line of research considering specular highlights as potential natural

adversaries and try to computationally model them, so that they can be taken care of. The related datasets, ImageNet-AH and ImageNet-PT, can be found at:
https://github.com/vanshikavats9/specular-adv

2. Related Work

2.1. Adversarial Attacks and Examples

Even the most powerful classifiers can be fooled into mis-predicting a class by introducing a simplest of aberration in the image captured. This was shown by Biggio et al. [4] using gradient based approach. Later, studies were demonstrated in formulating an adversarial attack with only a slight change in the pixel values [31], and then by fast gradient sign method [15]. Carlini et al. [6] introduce separate adversarial attacks on three different distance metrics. Along with these careful and optimised adversarial attacks, research has been done to generate real life examples of an adversary. Eykholt et al. [11] showed how the stop-signs classification gets affected by introducing small physical perturbations to the road signs which don't interfere with the human sign inference capability. Other studies show how simply digitally placing a random unrelated patch on the images can make it an adversarial image, thereby fooling the classifier [5], [32]. Hendrycks et al.[18] introduce a dataset of challenging images which are hard to classify by ImageNet pretrained classifiers even after being in the ImageNet-1K set. They adversarially select the samples which were fooled by ResNet-50 DNN model and also removing the samples which had a confidence of at least 15% in the correct class. This way, a carefully constructed dataset is empirically curated which gives very hard to predicted ImageNet samples. On the contrary, we present specular highlight as a natural phenomenon which can occur anywhere and anytime in the real world, affecting the classifiers' performance.

2.2. Specular Highlight

Specular highlight is a natural phenomenon which occurs as a bright spot on the objects when shined on by a light source. Its presence has proven to be a performance degrader in few of the computer vision tasks such as pixel-wise semantic image segmentation [1] and object detection [3], [13]. However, the effect of a specular highlight on image classification tasks has been underexplored. It becomes necessary to make our models robust to the glares and highlights which can affect our computer vision class prediction tasks in the real world scenarios, as mentioned in Sec.1. We thus aim to study how the specular highlight on the images degrades the performance of the DNN classifiers and give an insight about the need to address this robustness problem [18]. We show how the images were predicted correctly without the specular highlight, but as soon as we put a speckle on the images, the classifiers start to yield wrong predictions.

2.3. Class Activation Mapping

Class activation mapping is used to reveal the reason (model attention area) behind the deep learning model decisions. Grad-CAM [27] is one such method which uses the gradients of the targets to provide localized heatmap of the last convolutional layer. Grad-CAM++ [7] is an improvement on Grad-CAM analysis with better localization. They are now extensively being used in object localization and identification in combination with Mask-RCNN [22], and with other modifications such as Axiom-based grad-CAM by scaling gradients by normalized activations [14]. We use Grad-CAM++ due to its widespread and reliable use.

3. ImageNet-AH and ImageNet-PT

We present two datasets ImageNet-AH and ImageNet-PT, demonstrating the effect of specular highlight on the performance of the classifiers, using augmentation and torch flashing.

ImageNet Augmented Highlight (ImageNet-AH) is a collection of specular highlighted images (using augmentation) with specular highlights at various locations and intensities. Containing $\sim$142,000 adversarially affected images, this specially curated dataset contains those specular augmentations that fooled the classifiers into making a wrong decision but had their original images (I_o) correctly classified (say $I_{s,o}$). This will help us to claim that when we put a specular highlight on the image, the models change their decision towards a misclassification. The dataset is curated in such a way that we get a pool of failed images, $I_{s,o}$, mispredicted by any of the pre-trained classifiers considered in this study. The information of each failure (e.g. location and intensity of specular highlight) along with details of the network that misclassified it is stored in their meta-data. We select the same subset of 200 classes in ImageNet-1K pointed out by [18] such that the difference between them is prominent and the errors highlight the fault of the classifier. For example, a classifier wrongly classifying "ostrich" as a "porcupine" does more harm than it misclassifying "Norwich terriers" as "Norfolk terriers". Care is taken to avoid the rare classes such as "snow-leopard" and furthermore reduced overlapping classes like "honeycomb", "bee", "beehouse" etc. The full list of 200 classes spanning over broad categories is listed in the Supplementary Material.

ImageNet Print+Torch (ImageNet-PT) is a collection of 555 images with actual specular highlights (obtained by flashing physical torch) whose original image was correctly classified but one of the augmentations was wrongly classified by ResNet101. This helps to narrow down our selections. Since curating this dataset is more of a manual labor, a subset of 10 classes from ImageNet-AH is chosen to

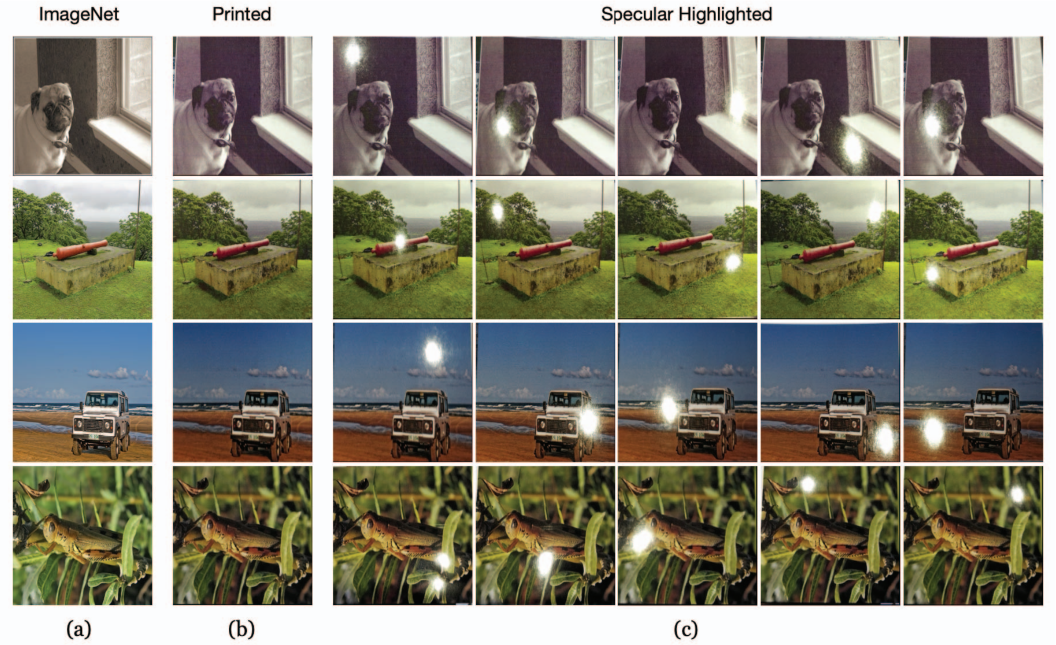

Figure 3. Specular Highlights on Natural Images to form ImageNet-PT. (a) 10 such classes used in ImageNet-AH are (b) printed out on a glossy paper and (c) specular highlights are formed using flashlight

make the experiment feasible as the samples were needed to be physically printed out on a glossy paper and actual light had to be thrown on them to make a speckle of highlight on the image. ResNet-101 is chosen because of its widespread use and good classifier performance on ImageNet-AH.

The performances are compared according to the Top-1 accuracy across the experiments between images with and without specular highlights. The analysis is also done to know which areas on the image contribute more to the performance degradation.

3.1. ImageNet-AH

To curate ImageNet-AH, each of the images in consideration was introduced to mathematically generated specular highlight to test the effect of the adversary. Each image was divided into a 5x5 grid and a highlight was introduced as a 2-dimension (2D) Gaussian distribution at the center of each of the grid cells (see Eq. 1). Here, $\forall\ x$, $y \in$ dim(image), we move the locations of the Gaussian kernel by shifting the center coordinates x_0 and y_0. The intensity of the highlight is varied with respect to $\sigma_x = \sigma_y = \sigma$ is the Standard Deviation (SD) of the distribution, varying as a step of 10 i.e. SD=10, 20, 30, 40 (Fig. 2). With 4 such unique SD intensities and 25 locations for specular highlights, we get 100 variations for each image. Top-1 accuracies are measured by comparing the original image dataset without the highlight and further quantifying how many specular variations of each image fail, differing in intensities and location.

$$f(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\left((x-x_0)^2 + (y-y_0)^2\right)\right) \tag{1}$$

The final dataset is curated in a way that the models make correct predictions on the original images without the specular highlight, but fail to classify into correct class when adversarially affected by the specular highlight.

3.2. ImageNet-PT

For ImageNet-PT, a subset of 10 image classes from ImageNet-AH, each containing about 10 images per class, is printed on glossy photography papers. The images are such selected that they are originally classified correctly by ResNet-101 in Experiment-I but their specular highlight variation failed. This way, the current experiment would be able to prove the claim made by the previous experiment about the failure classifiers in the case of specularly affected adversarial images. The aim is to show that the natural images are adversarially affected by an actual specular reflection and can fool the classifiers into misclassification. Various alterations of natural specular highlights are thrown on the printed images via a smartphone's flashlight to mimic an actual scenario of the objects/images affected by these adversaries (Fig. 3). The accuracies of with and without specularly affected images are reported in a similar manner as the mathematically generated highlights.

3.3. Heatmap analysis

Class activation mapping using Grad-CAM++ is used in this study to analyse which areas are the models interested in looking at to make their prediction. Grad-CAM++ is an improved version of Grad-CAM which uses gradient weighted class activation mapping looking at the last convolutional layer of a deep neural network to produce localization maps highlighting the attention regions responsible for making the class predictions. The heatmaps are generated for both the original images and their specular counterparts to see if the focus areas for both lie in the same location or are they also affected by the presence of a specular highlight.

4. Experiments

In this study, we emphasize the effect of specular highlights on eight deep learning models known for their image classification prowess, namely VGG16[29], VGG19[29], ResNet-50[16], ResNet-101[16], DenseNet-121[20], InceptionNet-v3[30], MobileNet-v2[26], and Xception[8]. The study spans over two variations of the specular adversary - mathematically created specular highlights and actual highlights on natural images. With each of the variations, we try to show how a small light spot on the images can easily fool a classifier even though they are easily distinguishable by the human eye.

4.1. Experiment I - Augmented Specular Highlights

A subset of 200 classes is taken out of ImageNet-1K's 1000 classes according to Hendrycks et al. [18] ImageNet-1K's validation set is chosen for this experiment as the images are available along with their class labels. Each class has 50 images of various dimensions, amounting to a total of 10,000 original images.

To make a simulated specular highlight on images, the images are divided into a 5x5 grid and a Gaussian noise with varying SD is introduced in each of the 25 grid cell locations (Eq. 1). SD=10, 20, 30, and 40 are taken into consideration for each grid cell location, making it a total of 100 specular variations per image. The top-1 accuracies of the model predictions of the original images are averaged over the classes as shown. For specular adversary performance on the affected images, the accuracies are averaged out from the 100 variations per image. Analysis is also done with respect to the instances where even 1 out of 100 variations fail and where at least 5 variations fail to consider a model prediction failed, both proving that this specular highlight works as an adversary to the image. Finally, the specular images mis-predicted by the models with their original image counterparts predicted correctly are curated in ImageNet-AH. As per the statistical analysis, 15.6% of the images had specular highlights of SD = 10, 21.0% had SD = 20, 27.7% had SD = 30 and the most (35.6%) had SD = 40. As we can see, the proportion increases with the SD, which is expected as larger portions in the images are getting blocked by the highlight.

4.2. Experiment II - Actual Highlights on Natural Images

To support the claim made by the synthetic highlights in Experiment-I, the images are tested with natural specular highlights. The images are printed out on a glossy photography paper and actual specular reflections are made by throwing a smartphone's flashlight on them. The images with this actual highlight is clicked by a camera and tested on the classifiers. To make this experiment feasible, a subset of 10 classes is selected to be printed from the original 200 classes. 5 variants of specular highlight at different locations are thrown on them. The performance is measured in a similar way as Experiment-I. Analysis is also done by thresholding the failure of the model prediction with at least 2 variations mispredicted.

4.3. Experiment III - Fine-tuning with Specular Adversaries

Original images correctly detected by each classifier but failing in their specular counterparts in Experiment I are saved and are used to fine-tune the default classification models to see if the performance can be improved. Training, validation and test sets from the 200 classes are separately formed for each classifier according to the number of images failed by each in a ratio of 60:20:20. Instead of just randomly dividing the specularly affected images into three sets, the division is done in a stratified manner. Care is taken to put all specular variations of a single original image into one set. This way, one kind of sample variations in one set would be unseen in the other sets giving a better intuition into the performance of the models.

Performance is also checked with finetuning on the datasets normalised with their respective means and standard deviations per channel. Further, we introduce self-attention into the networks using Squeeze-and-Excitation (SE) mechanism [19] to leverage the inter-channel dependencies and increase their representational power. The performance on the test set is compared on the various methods of finetuning used and contrasted with using the networks directly without finetuning.

4.4. Experiment IV - Activation Mapping

We study the class activation maps of both original and adversarially affected images to compare how the models are perceiving the information available on the images. We use Grad-CAM++ [7] and analyze where and what the models are looking for to make a prediction. Grad-CAM++ outputs the heatmap of attention areas in each image. To sepa-

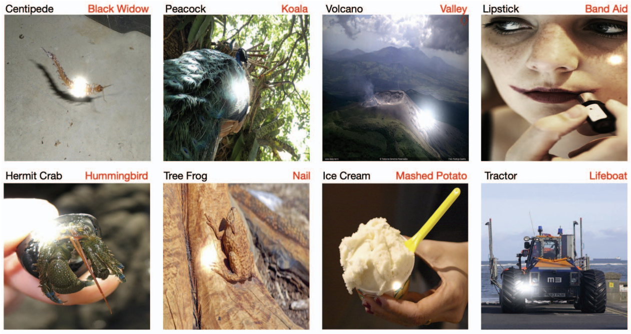

Figure 4. Examples of misclassification on specular highlight affected images on artificial specular highlights (Exp.I). The actual class is represented by black text and the incorrect prediction is represented by red text on the top of each image

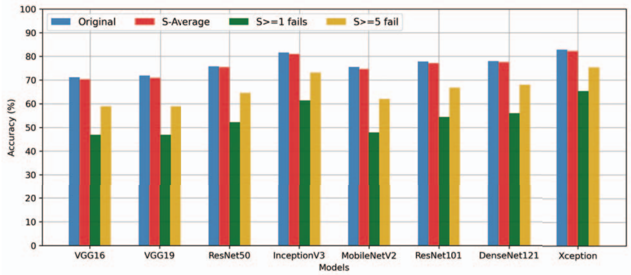

Figure 5. Performance of models for predicting original and artificial specular highlight affected images as the top-1 accuracy. We consider two scenarios: a model is considered failed if (i) it misclassifies even one specular variation (S>=1) of the original image or (ii) it wrongly predicts at least 5 specular variations (S>=5)

rate the most sought areas as determined by Grad-CAM++, binary thresholding is applied on the heatmaps with a cutoff at the third quartile (Q3) value of the pixel values in the heatmap (see Fig. 9). This gives an approximate area of which information contributes the most to the model prediction. The same procedure is followed for both original and specularly affected images in ImageNet-AH (Fig. 9(a)) and ImageNet-PT (Fig. 9(b)).

The thresholded heatmaps are divided into 5x5 grids and the cells with the maximum non-zero pixel (max-cells) values are determined. The max-cells of the original images are compared with the max-cells of the specular images. Specular variations for each original image are such chosen that they had earlier failed the prediction test in Experiment-I and II. This helps us to analyze how different each model is looking at the original and its specular highlight variation. Percentage of the max-cells of specular variations failing to match the max-cell of the original images per class is calculated for each model which indicates that the specular highlights cause the models to focus on a different location as compared to original distribution which might be the reason for their prediction failure.

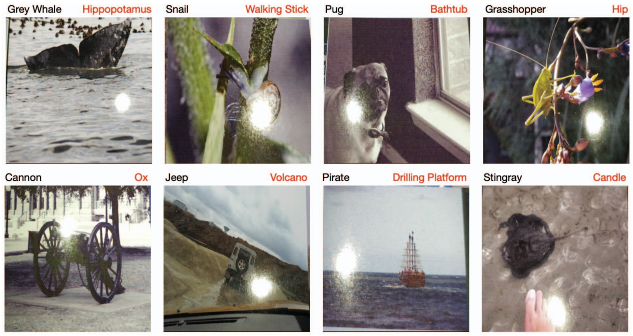

Figure 6. Examples of misclassification on specular highlight affected images on natural specular highlight (Exp. II). The actual class is represented by black text and the incorrect prediction is represented by red text on the top of each image

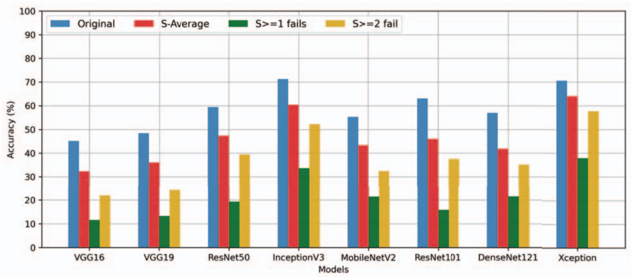

Figure 7. Performance of models on natural specular highlight affected images. We consider two scenarios: a model is considered failed if (i) it misclassifies even one specular variation (S>=1) of the original image or (ii) it wrongly predicts at least 2 specular variations (S>=2)

5. Results

The performances on the original and specular highlight adversarially affected images are compared. Fig. 4 shows some of the misclassifications due to the effect of a specular highlight to form ImageNet-AH. It can be observed that it's possible that a small highlight anywhere on the image can fool a classifier into making a wrong prediction. A detailed analysis is shown in Fig. 5 where we compare each model with respect to the Top-1 accuracy for original and their specular variations. An round figure of ∼20% drop in the accuracy is observed if we consider the model fooled when at least one specular variation gets misclassified (S>=1). There is ∼10% drop in the accuracy when a model is considered fooled if it misclassifies at least 5 or more specular variations (S>=5). Note only a slight effect in the average accuracy of the specular highlight variation failure. This is because of the average calculation strategy in case of 100 specular variations for each original image. For instance, if a single original image gets misclassified by a classifier, it is either 100% correctly classified or 0% correctly classified (wrong prediction). However if, out of the 100 specular highlighted variations of a wrongly predicted original image, even five get correctly predicted by chance, the accuracy will be 0.05% which is more than the original single image accuracy of 0%. Thus, S-average gives a lesser idea of the actual effect of a specular adversary.

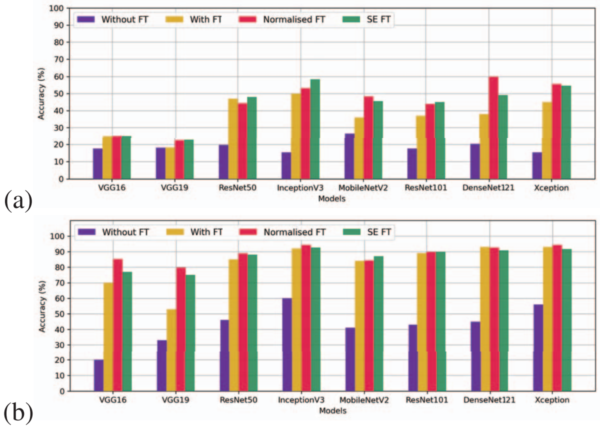

Figure 8. Effect of Fine-Tuning (FT) the models on the (a) augmented specular highlight images from ImageNet-AH and (b) natural specular highlighted images from ImageNet-PT. Analysis is also done with normalizing by specific means and standard deviations and by introducing squeeze-and-excitation (SE) block to the networks.

A similar pattern is observed when we test the models on the natural specular highlights on the printed images in ImageNet-PT (see Fig. 6, Fig. 7). The already low accuracy on the original images can be a result of different lighting conditions, different angles of capture, or just due to the printing of the digital image as mentioned in Kurakin et al. [24]. There is ∼12% drop is observed in the average accuracy, and a ∼35% and ∼20% drop if at least one variation fails (S>=1) and at least two variations fail (S>=2) respectively. Fine-tuning on the test split of the artificial specular highlight image datasets does not show enough improvement, averaged at 18% (Fig. 8(a)). Only InceptionV3 and Xception models show an improvement of >30% with fine tuning. Some improvement is seen when finetuned on the natural specular highlighted images (Fig. 8(b)). However, this might not be a true concluding insight because of a smaller test set of only 109 samples.

Normalizing the images by means and standard deviations of the respective datasets for each channel and then finetuning increases the performance by more than ∼5% compared to simple finetuning for all classifiers other than VGG variations and ResNet-50 for ImageNet-AH and significantly in VGG-16 and VGG-19 in ImageNet-PT. On the other hand, introducing the SE self-attention blocks in the fine-tuning layers does not give a clear trend in the improvement or the degradation of performance making it classifier-specific.

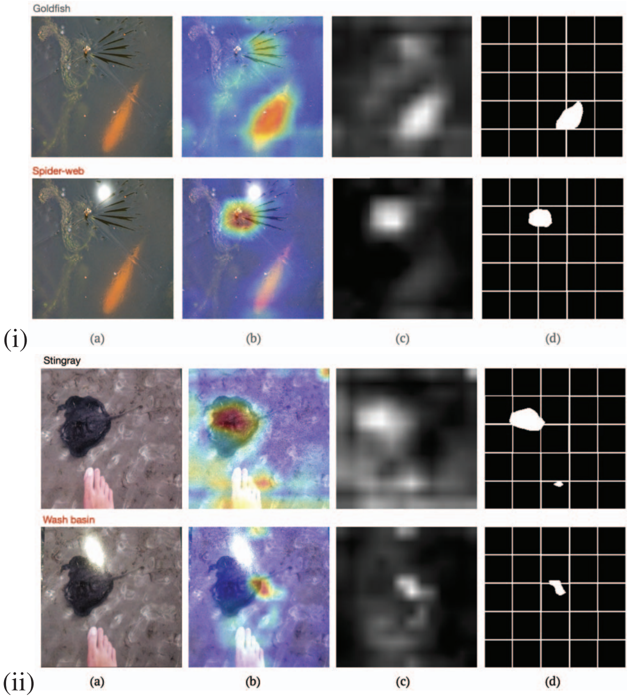

Figure 9. Activation analysis on the (a) original images (top row) of (i) ImageNet-AH and (ii) ImageNet-PT and their specular counterpart (bottom row). (b) Images superimposed with their (c) heatmaps depicting the most focussed areas. Heatmaps are divided into a (d) 5x5 grid and binary thresholded at Q3. The location of the max-cells are compared for the original and their specular counterparts. Here, it is seen that the specular adversary made the model concentrate on the wrong object and might contribute in making a wrong prediction (in red)

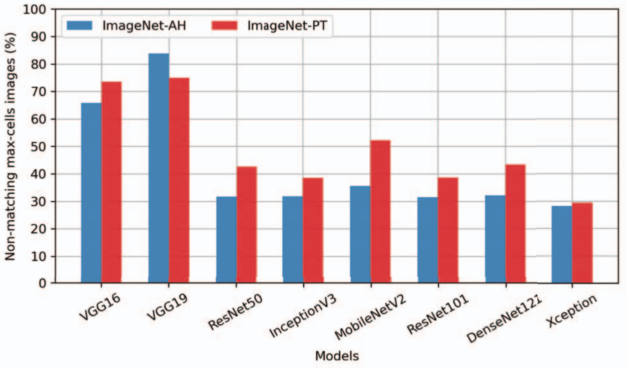

Figure 10. Class activation mapping analysis on the original and specular images in ImageNet-AH and ImageNet-PT. The bars indicate the percentages of sets of original-specular images whose attention location areas do not match

Activation analysis is also performed to look at where the models are paying attention to in case of the original images and their specular variations in ImageNet-AH (Fig. 9(a)) and in ImageNet-PT (Fig. 9(b)). It is observed that more than at least 65% of the VGG variations and ∼30% for rest of the models, the comparison original-specular sets for each model classifier exhibit a different location of focus

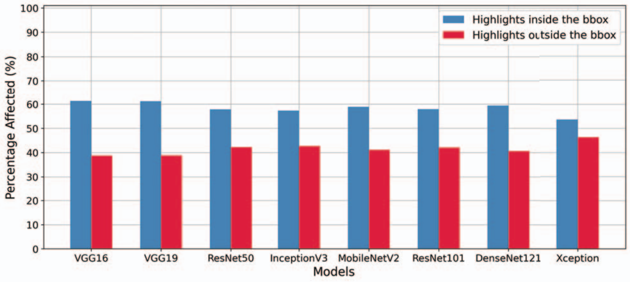

Figure 11. Percentage of highlights directly affecting the foreground object (inside the bounding box) or the background (outside the bounding box).

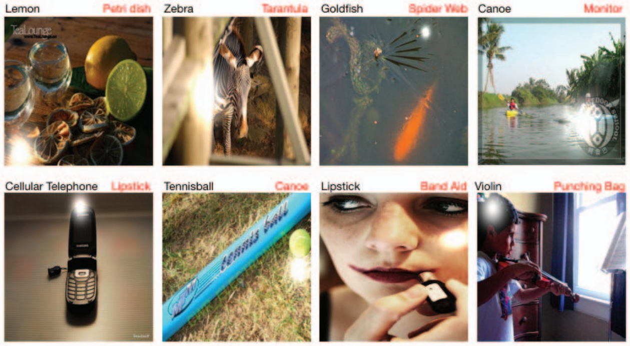

Figure 12. Examples of misclassification on the affected images when specular highlight is not directly on the object. The actual class is represented by black text and the incorrect prediction is represented by red text on the top of each image

areas (Fig. 10) for ImageNet-AH. This might be one of the reasons the classifiers make a different prediction when it sees a normal original image vs when we've put a specular highlight on it.

Since the ImageNet dataset also contains bounding box annotations for object detection, we make use of them to analyse whether the specular highlight put on the object foreground or the background. This is done to check if only the prominent objects in the images occluded due to the highlight are causing the mispredictions or the images are being affected even from the highlight far from the class deciding object. As observed from the statistics from all the models (Fig. 11), an average 59% of the highlights were inside the object bounding boxes, meaning, 59% of images had their main class objects being partially or fully occluded for the models to make their mispredictions. However, a massive 41% of the mispredicted images had their highlights far away from the object. This is a huge percentage and a matter of great concern that the images are being specularly affected even though the highlight is not directly on the class-defining object but on the surroundings. Fig. 12 shows some examples of the images being mispredicted with the specular highlights not being directly on the object. Also, on an average, 25 variations of the original images failed for all the models.

Inspection is also done to look for the location where a specular highlight put on the image would affect the prediction performance of the classifier the most. It is observed that a speckle on and near the center of the image will make the most images make a wrong prediction (Fig. 13). Since, most of the object instances are situated near the centre of the images, hence, a specular highlight present in that region can affect the prediction performance adversarially. Nevertheless, the above results indicate that specular highlight has the potential to act as an adversary and fool the classifiers into making the wrong predictions about the image class.

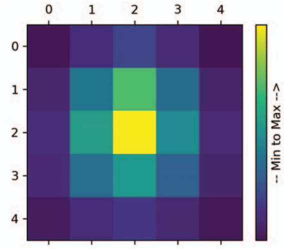

Figure 13. Heatmap depicting the locations in a 5x5 image grid where a specular highlight affects the image. Most adversarially affected images with wrong predictions have a specular highlight at their center, followed by the surrounding neighbourhood.

6. Conclusion

In this study, we show that the natural phenomenon of specular highlight on an image can adversarially affect the performance of modern-day classifiers. We demonstrate this by introducing two new datasets: ImageNet-AH, which is a mathematically induced specular highlight dataset, and ImageNet-PT, a dataset curated by flashing natural highlight on the images. We also see how specular highlights can shift the attention significantly. To that end, finetuning our models with various adaptations of the specularly affected images and networks could also not improve the performance enough with respect to what is desired. This underscores the importance of addressing the issue to improve the robustness of the models in real-world scenarios. Future studies will be directed towards exploring mitigation strategies to overcome the challenges posed by specularly affected images.

References

[1] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011. 1, 3

[2] Alessandro Artusi, Francesco Banterle, and Dmitry Chetverikov. A survey of specularity removal methods. *Comput. Graph. Forum*, 30:2208–2230, 12 2011. 2

[3] Shida Beigpour and Joost van de Weijer. Object recoloring based on intrinsic image estimation. In *2011 International Conference on Computer Vision*, pages 327–334, 2011. 1, 3

[4] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 387–402, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. 3

[5] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *ArXiv*, abs/1712.09665, 2017. 1, 3

[6] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, Los Alamitos, CA, USA, may 2017. IEEE Computer Society. 3

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 3, 5

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions, 2017. 5

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[10] Mahdi Abolfazli Esfahani and Han Wang. Robust glare detection: Review, analysis, and dataset release. *CoRR*, abs/2110.06006, 2021. 2

[11] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. 1, 3

[12] Yuming Fang, Zhenzhong Chen, Weisi Lin, and Chia-Wen Lin. Saliency-based image retargeting in the compressed domain. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, page 1049–1052, New York, NY, USA, 2011. Association for Computing Machinery. 2

[13] Gang Fu, Qing Zhang, Qifeng Lin, Lei Zhu, and Chunxia Xiao. *Learning to Detect Specular Highlights from Real-World Images*, page 1873–1881. Association for Computing Machinery, New York, NY, USA, 2020. 3

[14] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *CoRR*, abs/2008.02312, 2020. 3

[15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1, 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 5

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 1

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, June 2021. 2, 3, 5

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 5

[20] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. 5

[21] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3296–3297, 2017. 1

[22] Xavier Alphonse Inbaraj, Charlyn Villavicencio, Julio Jerison Macrohon, Jyh-Horng Jeng, and Jer-Guang Hsieh. Object identification and localization using grad-cam++ with mask regional convolution neural network. *Electronics*, 10(13), 2021. 3

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1

[24] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017. 1, 7

[25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 5

[27] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 3

[28] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 5

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. 5

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. Jan. 2014. 2nd International Conference on Learning Representations, ICLR 2014 ; Conference date: 14-04-2014 Through 16-04-2014. 1, 3

[32] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 49–55, 2019. 3

[33] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3395–3402, 2015. 2

[34] Y. Zhong, X. Liu, D. Zhai, J. Jiang, and X. Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15324–15333, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. 1