Using Image Quality Metrics to Identify Adversarial Imagery for Deep Learning Networks

Josh Harguess, Jeremy Miclat, Julian Raheema
Space and Naval Warfare Systems Center Pacific
53560 Hull Street, San Diego, CA 92152-5001
joshua.harguess@navy.mil, jeremy.miclat@gmail.com, julian.raheema@navy.mil

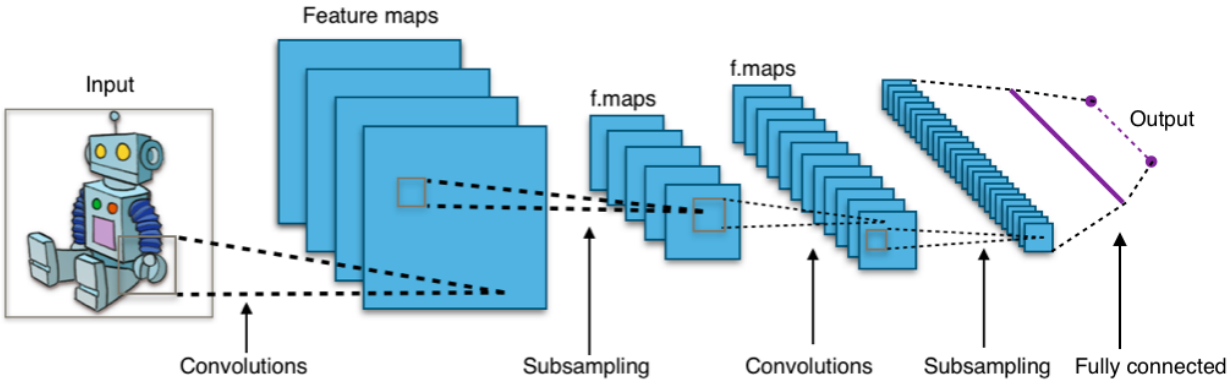
ABSTRACT

Deep learning has continued to gain momentum in applications across many critical areas of research in computer vision and machine learning. In particular, deep learning networks have had much success in image classification, especially when training data is abundantly available, as is the case with the ImageNet project. However, several researchers have exposed potential vulnerabilities of these networks to carefully crafted adversarial imagery. Additionally, researchers have shown the sensitivity of these networks to some types of noise and distortion. In this paper, we investigate the use of no-reference image quality metrics to identify adversarial imagery and images of poor quality that could potentially fool a deep learning network or dramatically reduce its accuracy. Results are shown on several adversarial image databases with comparisons to popular image classification databases.

1. INTRODUCTION

In very recent years, deep learning methods, which are algorithms and architectures built from multiple layered artificial neural networks (ANN), have gained tremendous momentum in solving many problems in machine learning (ML) and artificial intelligence (AI). In particular, the computer vision research areas of image classification and object detection have benefited greatly from deep learning methods, partly due to the advent of the convolutional neural network (CNN), which is a particular flavor of deep learning methods which takes inspiration from biological vision.[1–3] An example architecture for a typical CNN is shown in Figure 1.

Deep learning methods, such as CNNs, can solve very difficult high-dimensional machine learning problems, such as classification, using a large amount of labeled (annotated) data from the target problem. Once the weights of the network are initialized, the labeled data are fed forward through the network and new weights of the neurons in the layers are learned based on methods such as a back-propagation to reduce the error of the output of the deep learning network. This approach to machine learning has led to major breakthroughs in several problems, such as the ImageNet challenge,[1,5] and have led to an enormous increase in exposure to the progress in the area of machine learning from media outlets across the world.



(a) Example image from RainDropsRelax video
Figure 1. Typical architecture for a CNN.[4]

However, several researchers have begun investigating vulnerabilities with deep learning systems. Recently, a study by Nguyen, Yosinski, and Clune[6] demonstrated that these methods can be easily fooled by images that look nothing like images from the labeled imagery of the same category to the human eye. In their study, the "fooling" images are made by either directly or indirectly evolving imagery until the network believes with 99% confidence that the image belongs to a particular category of the classification problem. For example, this leads to images of TV static being classified as a strawberry (in the direct encoded case, as in Figure 2b). We denote these types of imagery that are capable of fooling a deep learning method as "adversarial" imagery.

Another growing area of research within computer vision is that of no-reference image quality metrics. No-reference metrics assume that no original or pristine example of what the image should look like is present at the time of assigning a metric to the quality of a given image. The blind/reference-less image spatial quality evaluator (BRISQUE) image quality metric introduced by Mittal et al.[7] has proven to be quite popular in many image quality applications.

In this paper we study the efficacy of utilizing no-reference image quality metrics, such as BRISQUE, for identifying adversarial imagery before they are presented to the deep learning network. The organization of the paper is as follows. In Section 2, we briefly describe the types of adversarial imagery[6] of interest to this paper. Then we briefly introduce the BRISQUE algorithm in Section 3. Section 4 introduces the dataset used for the study, describes our methodology, and gives the result of the study. We conclude the paper in Section 5.

2. ADVERSARIAL IMAGERY

We define adversarial imagery as imagery that intends (and often succeeds) to attack or fool an imagery recognition or classification system, such as a deep learning network. These types of imagery are more intentional and potentially more harmful than simple obfuscations or occlusions of an image to produce an incorrect classification. Instead, they are generated with the intent of obtaining a very high confidence of classification, even though to the human observer they are clearly not from the class for which they are classified.

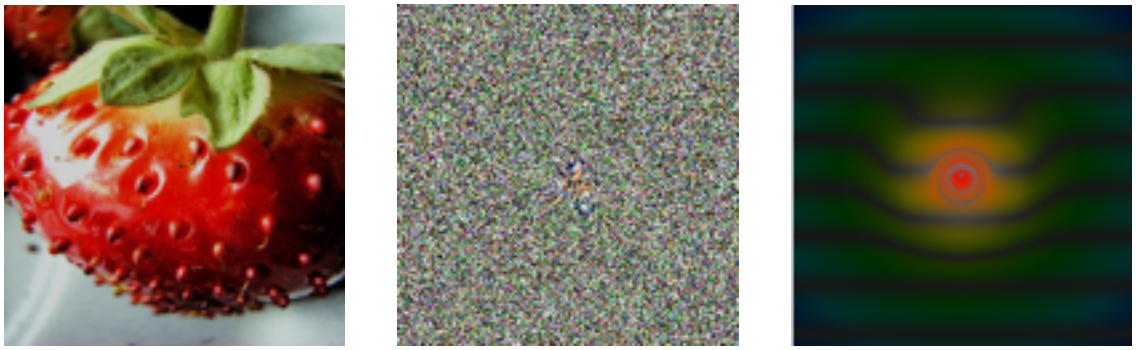
In this work, we study adversarial images, or "fooling" images, which are formed from evolutionary algorithms, such as those found in the work by Nguyen et al.[6] One type of adversarial image is generated from "direct encoding", meaning that the pixels themselves are mutated over many evolutionary steps to produce adversarial imagery. These direct encoded images are initialized with uniform random noise and then each pixel is independently mutated until the desired classification result is obtained. Examples of direct encoded images for the ImageNet category "strawberry" and for the MNIST handwritten digit 2 are shown in Figure 2.

Another type of adversarial image is generated from "indirect encoding", which uses a compositional pattern-producing network (CPPN)[8] to evolve complex images that can resemble natural and man-made objects. Examples of indirect encoded images for the ImageNet category "strawberry" and for the MNIST handwritten digit 2 are shown in Figure 2.

3. BRISQUE: NO-REFERENCE IMAGE QUALITY

No-reference image quality metrics are gaining in popularity for computing the "quality" of an image when no original image is present for comparison. Blind/reference-less image spatial quality evaluator (BRISQUE)[7] is an algorithm that uses natural scene statistics (NSS) to build a distortion-generic no-reference metric for image quality. The area of research within NSS studies the statistical regularities that are present within "natural scene" images and identifies and defines metrics for measuring how much the statistical properties of an unknown image vary from the statistics of usual natural scene images. For instance, if the unknown image is highly distorted, we would not expect the statistical regularity of that image to match that of usual natural images. Within the BRISQUE algorithm, mean subtracted contrast normalized (MSCN) coefficients are computed as a way to compare distributions of high quality natural images to test images. Examples of these MSCN coefficients of shown in Figure 3 produced from an image from the RainDropsRelax video[9] and an image from a low quality video of ships from an unmanned aerial vehicle (UAV).[10]

Small BRISQUE values indicate low distortion, and therefore high quality, while large values indicate high distortion and low quality.



(a) Strawberry (b) Direct encoded (c) Indirect encoded

(d) Handwritten digit 2 (e) Direct encoded (f) Indirect encoded

Figure 2. Example images from ImageNet (a) ("strawberry" category) and from MNIST (d) (handwritten digit "2") with corresponding direct and indirect encoded examples of the same category.

4. RESULTS

In this section, we introduce the datasets used for this paper, briefly explain the methodology for identifying adversarial imagery, and present our results.

4.1 DATASET

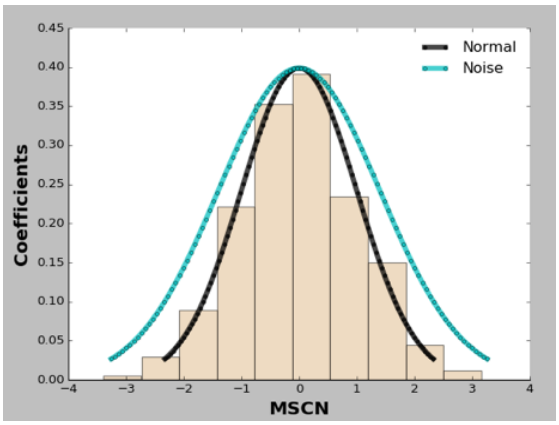
Our study uses two image sets (ImageNet and MNIST) for analysis. We have three groups for ImageNet dataset: original, direct encoded, and indirect encoded. Using BRISQUE, we calculate image quality for each group then compute the mean and standard deviation. We continue generating the mean and standard deviation for the direct and indirect encoded groups too. Finally we compute the Gaussian distribution for the three groups (Figure 4). For MNIST dataset, the same steps were applied to generate the distribution graphs for each of the three sets: original, direct encoded, and indirect encoded (Figure 4).

4.2 Methodology

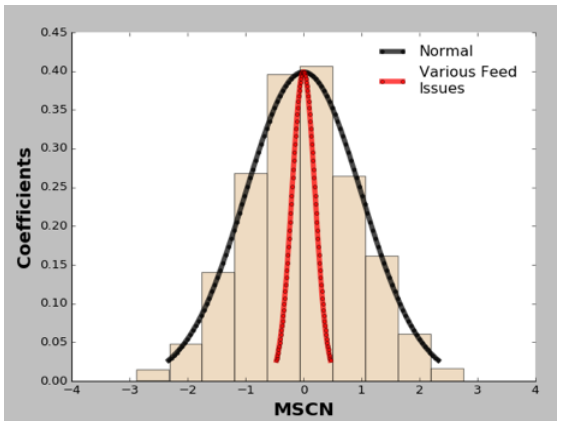
The methodology for identifying adversarial imagery in this work involves two steps. First we compute the BRISQUE scores for all original images and adversarial images (direct and indirect encoded in this case). These scores can be visualized in Figure 4 Second, we compute receiver operating characteristic (ROC) curves to illustrate the performance of our binary classifier (non-adversarial vs. adversarial) as the discrimination threshold is varied in Figures 4a and 4b using the BRISQUE score as the discriminating feature. The resulting ROC curves for ImageNet and MNIST (original vs. direct and original vs. indirect), respectively, are shown in Figures 5 and 6.

(a) Example image from RainDropsRelax video[9]

(b) Example image from UAV video[10]



(c) MSCN coefficients for RainDropsRelax image



(d) MSCN coefficients for UAV image

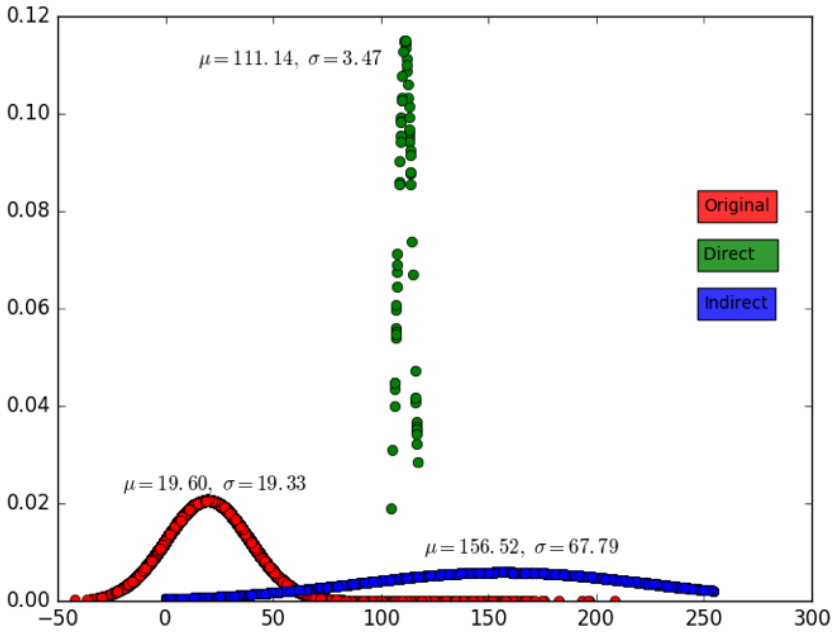
Figure 3. Example images and their MSCN coefficients for the BRISQUE algorithm.

4.3 Results

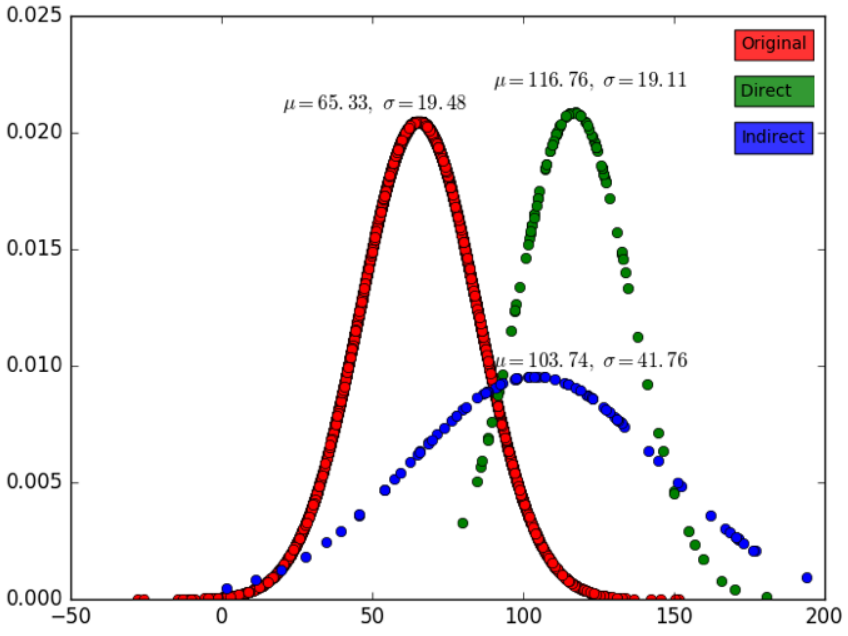
As we can see from the figures, the ImageNet data is quite separable from its direct and indirect encoded images, so this method of using BRISQUE for adversarial imagery appears to have good success, especially in the case of direct encoded imagery. However, MNIST and its adversarial imagery are more difficult to tell apart. This could be due to the fact that the BRISQUE score is trained on natural imagery, not handwritten digits.

5. CONCLUSION AND FUTURE WORK

In this paper we have introduced a preliminary study on using image quality metrics to identify adversarial imagery before the imagery reaches the deep neural network. The approach of using the BRISQUE algorithm to identify adversarial imagery has been shown to work quite well for ImageNet data, but not as well as the MNIST data. This is mostly due to the fact that BRISQUE is trained on natural imagery, not handwritten digits. We will extend this work to include results of BRISQUE trained on handwritten digits and other datasets focused on a particular classification problem. One area of future work is to study the usefulness of image quality metrics on newer adversarial approaches, such as that found with "universal adversarial pertubations"[11] for general object recognition and the use of facial attributes for face recognition.[12] In general, we hope to work towards the goal of building architectures that are robust to the adversarial imagery described in this paper as well as other potential adversarial imagery.

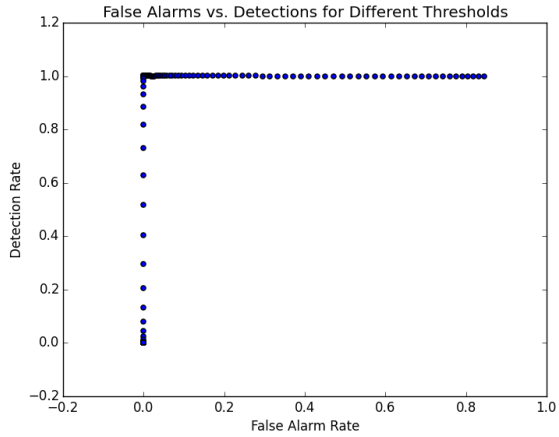


(a) BRISQUE scores for ImageNet

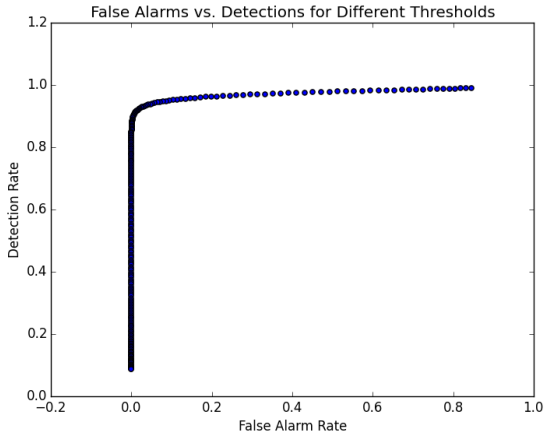


(b) BRISQUE scores for MNIST

Figure 4. BRISQUE score distributions for the ImageNet and MNIST datasets for the original (red) and the direct (green), and indirect (blue) encoded images.

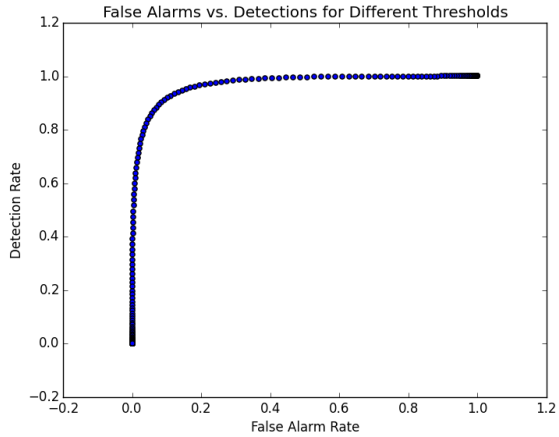


(a) ROC for ImageNet - Original vs. Direct

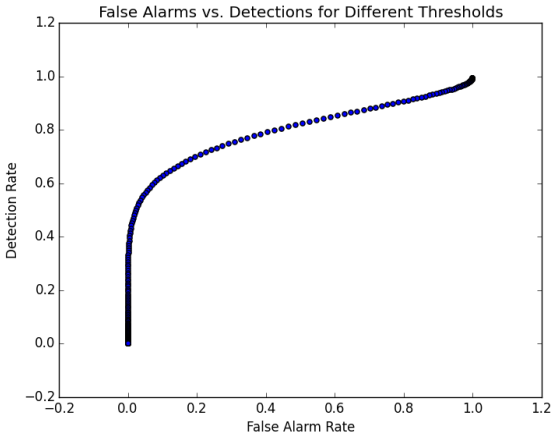


(b) ROC for ImageNet - Original vs. Indirect

Figure 5. ROC curves for the original vs. direct and the original vs. indirect cases for the ImageNet data.



(a) ROC for MNIST - Original vs. Direct



(b) ROC for MNIST - Original vs. Indirect

Figure 6. ROC curves for the original vs. direct and the original vs. indirect cases for the MNIST data.

REFERENCES

[1] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [Advances in neural information processing systems], 1097–1105 (2012).

[2] Hubel, D. H. and Wiesel, T. N., "Receptive fields and functional architecture of monkey striate cortex," The Journal of physiology 195(1), 215–243 (1968).

[3] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., "Gradient-based learning applied to document recognition," Proceedings of the IEEE 86(11), 2278–2324 (1998).

[4] Aphex34, "Typical cnn by aphex34 (own work) [cc by-sa 4.0 (http://creativecommons.org/licenses/by-sa/4.0)], via wikimedia commons." `https://upload.wikimedia.org/wikipedia/commons/6/63/Typical_cnn.png` (December 2015).

[5] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., "Imagenet large scale visual recognition challenge," International Journal of Computer Vision 115(3), 211–252 (2015).

[6] Nguyen, A., Yosinski, J., and Clune, J., "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in [Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition], 427–436 (2015).

[7] Mittal, A., Moorthy, A. K., and Bovik, A. C., "No-reference image quality assessment in the spatial domain," IEEE Transactions on Image Processing 21(12), 4695–4708 (2012).

[8] Secretan, J., Beato, N., D Ambrosio, D. B., Rodriguez, A., Campbell, A., and Stanley, K. O., "Picbreeder: evolving pictures collaboratively online," in [Proceedings of the SIGCHI Conference on Human Factors in Computing Systems], 1759–1768, ACM (2008).

[9] Harguess, J., "Automated content and quality assessment of full-motion-video for the generation of meta data," in [SPIE Defense+ Security], 94630L–94630L, International Society for Optics and Photonics (2015).

[10] Parameswaran, S., Harguess, J., Barngrover, C., Shafer, S., and Reese, M., "Evaluation schemes for video and image anomaly detection algorithms," in [SPIE Defense+ Security], 98440D–98440D, International Society for Optics and Photonics (2016).

[11] Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P., "Universal adversarial perturbations," arXiv preprint arXiv:1610.08401 (2016).

[12] Rozsa, A., Günther, M., Rudd, E. M., and Boult, T. E., "Are facial attributes adversarially robust?," arXiv preprint arXiv:1605.05411 (2016).