Perceptual Evaluation of Adversarial Attacks for CNN-based Image Classification

Sid Ahmed Fezza[§], Yassine Bakhti[§*], Wassim Hamidouche[*] and Olivier Déforges[*]
[§]National Institute of Telecommunications and ICT, Oran, Algeria
[*]Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France
sfezza@inttic.dz, firstname.lastname@insa-rennes.fr

***Abstract*—Deep neural networks (DNNs) have recently achieved state-of-the-art performance and provide significant progress in many machine learning tasks, such as image classification, speech processing, natural language processing, etc. However, recent studies have shown that DNNs are vulnerable to adversarial attacks. For instance, in the image classification domain, adding small imperceptible perturbations to the input image is sufficient to fool the DNN and to cause misclassification. The perturbed image, called *adversarial example*, should be visually as close as possible to the original image. However, all the works proposed in the literature for generating adversarial examples have used the L_p norms (L_0, L_2 and L_∞) as distance metrics to quantify the similarity between the original image and the adversarial example. Nonetheless, the L_p norms do not correlate with human judgment, making them not suitable to reliably assess the perceptual similarity/fidelity of adversarial examples. In this paper, we present a database for visual fidelity assessment of adversarial examples. We describe the creation of the database and evaluate the performance of fifteen state-of-the-art full-reference (FR) image fidelity assessment metrics that could substitute L_p norms. The database as well as subjective scores are publicly available to help designing new metrics for adversarial examples and to facilitate future research works.[1]**

***Index Terms*—deep neural network, adversarial attack, adversarial example, subjective evaluation, perturbation**

I. Introduction

One can only be impressed by the deep neural networks (DNNs) performance that are significantly superior to those achieved using conventional shallower networks. Taking advantage of the proliferation of large datasets in addition to the increase in computational power, the DNNs have shown a high efficiency in various difficult tasks such image classification [1], object detection [2], speech recognition [3] and natural language processing [4]. For instance, in the field of image recognition, the DNNs are able to recognize images with almost human precision, allowing them to be used in different sensitive applications such as autonomous cars, biometric, video surveillance, etc.

Despite state-of-the-art performance achieved by DNNs, it has been shown that they are vulnerable and unstable to adversarial attacks [5]. For instance, in the field of image classification, Szegedy *et al.* [6] was the first to show that small and almost imperceptible perturbations added to test images could lead DNN to misclassifying them. The perturbed images are called *adversarial examples*.

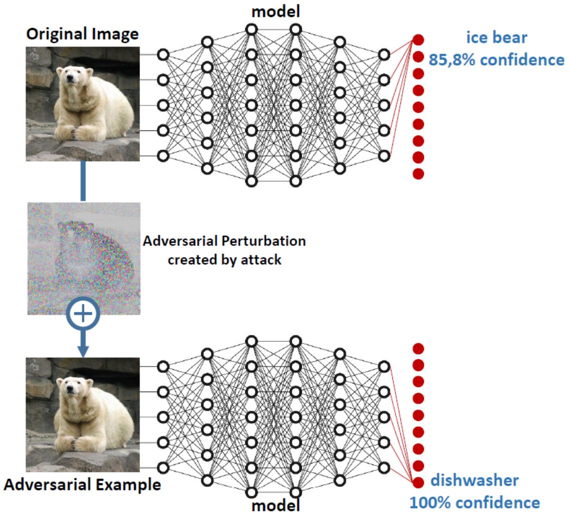


Fig. 1. Applying small imperceptible perturbation to the input image fool the deep neural network classifier. The original image is classified as an *ice bear* with 85.8% confidence, while the adversarial example is classified as a *dishwasher* with 100% confidence.

Goodfellow *et al.* [7] define adversarial examples as "inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake." Figure 1 shows how an original image carefully crafted using a small perturbation induces the network into misclassification with high confidence. Although, to a human, the adversarial image is indistinguishable from the original, *i.e.* the perturbation is quasi-imperceptible, the classifier labels them differently. This highlights the lack of robustness of the DNNs against adversarial examples, which raises security issues and limits the applications in which the neural networks can be deployed in a real-world environment. For instance, an adversary can use adversarial examples to manipulate the traffic signs so that the car takes undesirable and inappropriate actions, which is significantly dangerous. Therefore, it is of paramount importance to understand how and why these vulnerabilities to attacks occurs, thereby increasing the robustness of DNNs against the adversarial examples and bridging the gap between human perception and DNN-based systems.

Recently, great efforts have been made to propose methods for generating adversarial examples, which have been used as a benchmark for evaluating the robustness of candidate defenses.

[1] https://github.com/safezza/IQA-CNN-Adversarial-Attacks

Several adversarial attacks strategies have been proposed in the literature, and they are primarily differentiated by their computational cost, the level of knowledge about the attacked model and the purpose of the attacker [8]–[10].

As a general rule, two factors of adversarial examples, in the domain of image classification, should be considered [8]–[10]:

1) Fooling the image classifier: the adversarial example is crafted by adding small perturbation to the original image in such a way to cause classification mistake, *i.e.,* the perturbed image is misclassified to a specific class (*targeted attack*) or only misclassified to an arbitrary class (*untargeted attack*).
2) Imperceptible perturbation: the introduced perturbation should be undetectable by human observer. The original image and its intentionally perturbed version (adversarial example) are expected to be visually very close, and the differences between them are hardly noticeable by the human eye.

For the second aspect, which is the focus of our study, all the works proposed in the literature for generating adversarial examples have used the L_p norms (L_0, L_2 and L_∞) as distance metrics to quantify the similarity between the original image and the adversarial example [9], [10]. However, the L_p norms do not correlate with human judgment, because they are pixel-based error measures and do not take into account the properties of human visual system (HVS) [11]. Despite these common measures provide poor performance for assessing perceptual similarity/fidelity, all existing works have adopted these metrics as perturbation measures for generating adversarial examples.

On the other hand, in the last decade, considerable research efforts have been made to develop objective quality/fidelity assessment metrics [12]–[14]. The purpose of this research is to develop tools allowing to evaluate the quality/fidelity in a way that is consistent with human judgments [14]. There is a tendency to confuse image quality metrics with image fidelity metrics, despite the fact that they are closely linked, the two families of metrics have different purposes. The former are designed to predict subjective human appreciation upon the quality of multimedia content, while the latter refer to the ability to quantify the visual differences between a reference and test image [15]. Given the purpose of this work, image fidelity assessment (IFA) metrics are more appropriate for the generation and performance analysis of adversarial examples. That is why in the rest of this paper, we refer to image fidelity metrics instead image quality metrics.

However, the use of an inappropriate IFA metric can lead to wrong conclusions and suboptimal results, which can be the case with L_p norms that exhibit poor correlation with the human perception. There is therefore an urgent need to find a more accurate IFA metric that could substitute L_p norms for generating and assessing adversarial examples in close agreement with human similarity judgments.

The natural way to reach this goal is to take advantage of the many IFA metrics proposed in the literature. However, these metrics were typically developed for some specific applications, and consequently were designed to capture distortions that are related to these applications, such as blur and blocking for compression, noise for acquisition and fast fading for wireless transmission, to cite a few examples. Nevertheless, the adversarial perturbations/distortions used against DNNs can have different properties than those widely tackled by the quality/fidelity assessment community. Thus, developing new reliable IFA metrics specifically for adversarial examples represents a new research challenges to this community.

In this paper, we present a database for visual fidelity assessment of adversarial examples. To the best of our knowledge, this database is the first one specifically dedicated to the perceptual assessment of adversarial perturbations against DNNs and is publicly available to facilitate future research works. The dataset includes 360 images that have been generated using six prominent adversarial attacks with different levels of perturbations. The subjective data of eighteen human subjects have been collected, where each subject was asked to rate the fidelity of the adversarial example with respect to the reference image. The resulting MOS scores have been used to evaluate the performance of the three distance metrics (L_0, L_2 and L_∞) and to assess the performance of fifteen state-of-the-art full-reference (FR) image fidelity assessment metrics, as well as can be used to design new IFA metrics for adversarial examples.

The rest of this paper is organized as follows. Section II provides the taxonomy of adversarial attacks. Section III describes the performed subjective experiment, including the preparation of the test material, environmental setup and the test methodology. Next, the results and analysis of objective metrics are provided in Section IV. Finally, Section V concludes the paper.

II. Adversarial Attacks on Deep Neural Networks

An adversarial example is an original image carefully-crafted by an adversary attack with the aim to fool DNN classifier. The adversary attacks can be divided into two categories: *white-box attacks* that have a full access to the architecture and models parameters of the DNN, and those who only have access to the output of the attacked model (label or confidence score), known as *black-box attacks*. In addition, according to the objective to be reached, adversary attacks can also be distinguished as *targeted* and *untargeted* attacks. Formally, given an original input image x and a trained classifier C, generating an adversarial example x' can be formulated as a constrained optimization problem [9]:

$$\begin{aligned} x' = \ & \arg\min_{x'} \mathcal{D}(x, x'), \\ s.t. \ & C(x) = l, \\ & C(x') = l', \\ & l \neq l,' \end{aligned} \tag{1}$$

where $\mathcal{D}$ denotes a distance metric between two data sample, while l and l' denote the output class label of x and x',

TABLE I
SELECTED ADVERSARIAL ATTACKS PARAMETERS WITH THEIR CORRESPONDING VALUES.

Adversarial attack	Parameter	Values
FGSM [5]	ϵ	0.002, 0.03, 0.06, 0.14, 0.4
BIM [16]	ϵ	0.003, 0.03, 0.06, 0.15, 3
Deepfool [17]	overshoot	0.25, 1.0, 3.5, 36, 500
C&W [10]	(confidence, learning_rate)	(10, 0.4), (10, 1), (30, 0.9), (30, 1.3), (70, 0.9)
PGD [19]	ϵ	0.003, 0.03, 0.1, 0.4, 1.40
MIM [18]	ϵ	0.005, 0.03, 0.06, 0.19, 0.6

respectively. In the case of a target attack, l' is specified by the attacker, while for an untargeted attack, l' can be any class label, as long as it is different from the correct label l.

The distance metric $\mathcal{D}$ is used to quantify similarity/fidelity between the adversarial example and the original image. In the literature, three metrics are commonly used for generating adversarial examples, and all three are L_p norms [10]. In other words, the amount of perturbation is quantified by L_p norms, *i.e.,* $\|x - x'\|_p$, where the p-norm is defined as

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}} \tag{2}$$

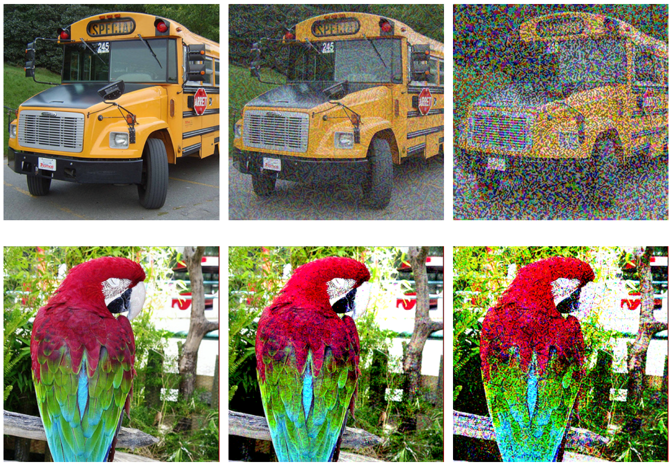
Specifically, L_0, L_2 and L_∞ are the three widely used metrics:

1) L_0 metric counts the number of pixels that have been altered in the adversarial example.
2) L_2 metric measures the Euclidean distance between the adversarial example and the original image.
3) L_∞ metric denotes the largest absolute difference value among all pixels in the adversarial example.

Nevertheless, these metrics do not correlate with human perception, because, they totally overlook spatial relationships between image pixels, and also consider that all changes in the visual signal are of equal importance. Finally, they do not take into account any of the perceptual properties of the HVS.

Several methods have been proposed in the literature to generate adversarial examples, they differ mainly in the modeling of objective function that seeks the best solution to the optimization problem described above in (1). The perturbation is determined by maximizing the classification error, while minimizing distance metric.

In our subjective study, we considered six prominent attacks that are: Fast Gradient Sign Method (FGSM) [5], Basic Iterative Method (BIM) [16], Deepfool [17], Carlini-Wagner (C&W) attack [10], Projected Gradient Descent (PGD) [19] and Momentum Iterative Method (MIM) [18]. All these attacks are gradient-based adversarial generating approaches [9]. Specifically, the input image is perturbed according to the gradient of the loss function of the attacked DNN, where the perturbation magnitude gradually increases until the image is misclassified. For a complete description of these attacks, the reader is refereed to their original papers.



(a) Small impairment (b) Medium impairment (c) High impairment

Fig. 2. Adversarial examples with different levels of impairment generated using BIM (top) and C&W (bottom) attacks.

III. Subjective Evaluation

In this section, the conducted subjective study of adversarial examples is presented. Our goal is to use the ground truth obtained from human judgments to check the suitability of several state-of-the-art image fidelity assessment (IFA) metrics for adversarial examples, which can constitute viable alternatives to the three widely used distance metrics (L_0, L_2 and L_∞).

A. Adversarial Attacks Description

As mentioned previously, a total of six adversarial attacks have been employed to generate the adversarial examples. All these attacks have been implemented using *Cleverhans* software library [20], which provides standardized reference implementations of adversarial example generation techniques. Each attack can be tuned through a set of parameters, here, we are only focused on the ones controlling the magnitude of the perturbation introduced. Table I lists the parameters used to generate the different adversarial examples. The parameters values have been carefully chosen in a way to generate adversarial examples with a broad range of perturbations/distortions, thus covering the full range of subjective impairment scale, from imperceptible levels to high levels of impairment. Figure 2 shows some samples of adversarial images, in addition, the histogram of subjective scores for the entire dataset is illustrated in Figure 4.

As a victim DNN model, we used the well-known Inception v3 network [21], because it is pre-trained on *ImageNet* dataset [22] that we considered as a source image, as reported in Section III-B. Thus, the gradient of its loss function is exploited by the different attacks to compute the perturbation introduced to the original input image.

B. Dataset Preparation

Since our work deals with adversarial examples for DNN-based image classification, we focused our subjective experiment on the most widely used image classification dataset,

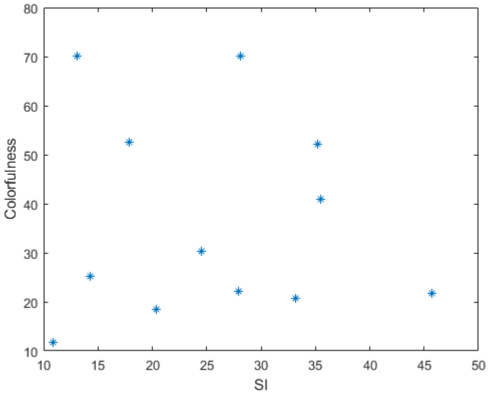


Fig. 3. SI and CF distributions of the selected contents.

which is *ImageNet* database [22]. Twelve images have been selected from the database that represent different content, including indoor and outdoor scenes and a wide range of colors and textures. In order to cover a wide range of features, the spatial complexity and color features of each image have been analyzed using Spatial Information (SI) [23] and ColorFulness (CF) [24], respectively. The Figure 3 shows the values of SI and CF for all the selected images.

The original images have different sizes, that we cropped to the size of 299×299 pixels covering the main object in the image. Because, given that we used Inception v3 network as attacked DNN, and the latter has an image input size of 299×299. Consequently, to avoid the up- and down- sampling operations that can introduce distortions to the input image, we made choice to crop the images to the input size of the Inception v3 network.

Thus, the twelve selected and cropped images were used to produce the subjective test dataset. Each image was perturbed/attacked using the six different adversarial attacks with the five different parameter settings, thus providing 360 adversarial examples. In addition, two other different images have been selected for training.

C. Environment Setup and Test Methodology

The subjective evaluations were conducted in a laboratory psychovisual test room, calibrated according to ITU-R BT.500-13 Recommendations [25], equipped with a controlled lighting system and the color of the all background walls and curtains is mid-gray. A full HD 27-inch Dell UltraSharp U2717D was used to display the test stimuli. The distance of the subjects from the monitor was approximately equal to 7 times the picture height, as recommended in [26].

Since the detection of impairment is an important factor in our study, the subjective experiments have been conducted using the Double Stimulus Impairment Scale (DSIS) method [25]. Both the original image and adversarial example were displayed in a side-by-side arrangement on the same monitor. The original image and adversarial example were always displayed on the left and right side, respectively, and the subjects were aware of these positions.

At the end of the presentation of each pair of images, a dedicated user interface was displayed on the screen for about five seconds during which the subject gives its judgment. The participants were asked to rate the level of impairment of the adversarial examples with respect to the reference original image, using a five-grade discrete impairment scale (1: very annoying, 2: annoying, 3: slightly annoying, 4: perceptible, but not annoying, 5: imperceptible). In other words, the observers tried to quantity the visibility degree of the perturbation introduced by the attack.

Given the large number of stimuli, making impossible to show all of them in a single session, because the viewing session would exceed 30 minutes. Consequently, in order to avoid visual fatigue effects, the subjective experiment was divided into three sessions whose duration does not exceed 20 minutes each. Subjects took a break between each two sessions. Moreover, each test session involved only one subject assessing the stimuli. In order to avoid possible contextual and memory effects, the display order of these stimuli was randomized in a way that the same content was never shown consecutively.

Before the experiment starts, instructions explaining the task were provided to subjects. In addition, training session was held with additional images, allowing the subjects to practice and become familiarize with the test procedure. The quality of these training samples was chosen so that it covers the full rating scale.

A total of 18 naive subjects (5 females and 13 males) took part in the subjective experiment. The age of subjects was ranging from 21 to 54, with an average of 28.8. All subjects were screened for color blindness and visual acuity using Ishihara and Snellen charts, respectively.

D. Data Processing

First, the subjective scores were processed to detect and exclude possible outliers, *i.e.,* subjects whose scores deviated strongly from others. Outliers detection was performed as specified in [25], and no outlier subjects were detected in this study.

Second, the Mean Opinion Score (MOS) was computed as the mean across scores provided by different subjects as follows:

$$MOS_j = \frac{1}{N} \sum_{i=1}^{N} s_{ij} \tag{3}$$

where N is the number of subjects and s_{ij} is the score given by subject i for the stimulus j.

In order to evaluate the reliability of the obtained results from statistical point of view, 95% confidence intervals (CI), assuming a Students *t*-distribution of the scores, were computed together with MOS values.

IV. Objective Evaluation and Results Analysis

It is highly desirable that the obtained MOS scores show fair distribution of values and are representative of the different impairment level on the rating scale. Figure 4 shows

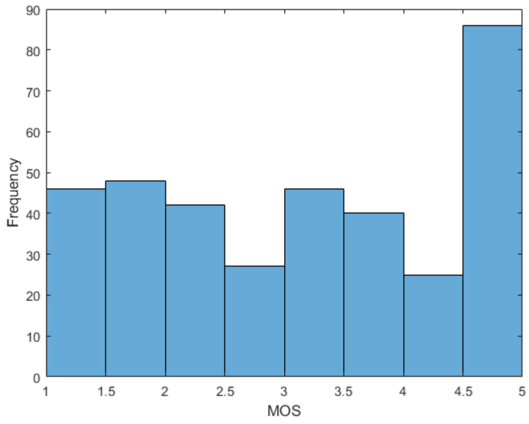


Fig. 4. Histogram of the MOS scores in the database.

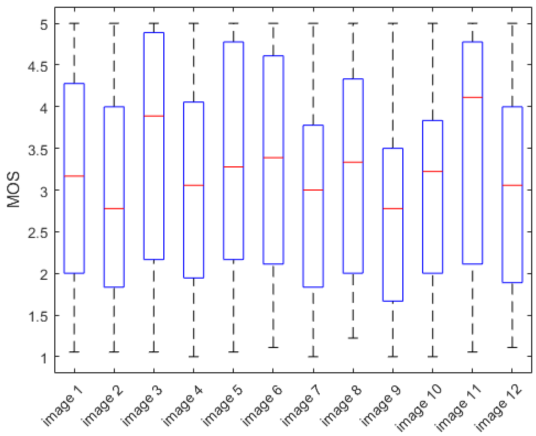


Fig. 5. The distribution of MOS values for each image.

MOS values distribution on the whole database. Overall, we obtained an almost a fair distribution, except for the 4.5-5 scale for which we obtained higher frequency. This mainly due to Deepfool attack, which impairment level is hardly to adjust and often provides undetectable perturbations. In addition, Figure 5 illustrates the distribution of MOS values for each assessed image. Thus, the resulting MOS values uniformly span the whole impairment scale, which means that the subjective experiments have been properly designed and conducted.

The results of the subjective tests were used as ground truth to evaluate fifteen full reference (FR) objective fidelity/quality metrics, namely: Peak-Signal-to-Noise-Ratio (PSNR), Structural Similarity Index (SSIM) [27], Feature Similarity Index (FSIM/FSIMc for color images) [28], Visual Signal-to-Noise Ratio (VSNR) [29], Gradient Similarity Measure (GSIM) [30], Most Apparent Distortion (MAD) [31], Multi-Scale SSIM index (MS-SSIM) [32], Visual Saliency-based Index (VSI) [33], Visual Information Fidelity (VIF/VIFp for pixel domain) [34], Information Fidelity Criterion (IFC) [35], Weighted Signal-to-Noise Ratio (WSNR), Universal Quality Index (UQI) [36], Noise Quality Measure (NQM) [37].

In addition, the three widely used distance metrics (L_0, L_2 and L_∞) have been also considered for evaluation and are compared against the fifteen objective fidelity metrics.

The performance evaluation of the set of metrics has been carried out in terms of three attributes: accuracy, monotonicity, and consistency, with respect to subjective scores. To achieve this goal, four performance measures were used, namely Pearson Linear Correlation Coefficient (PLCC) and Root Mean Square Error (RMSE) for prediction accuracy, while Spearman Rank Order Correlation Coefficient (SROCC) and Outlier Ratio (OR) for monotonicity and consistency, respectively. We can say that a metric obtains good performance, if the values of PLCC and SROCC are high (close to ± 1), and the values of RMSE and OR is low (close to 0).

PLCC measure was computed between the MOS and the objective score (Q_p) provided by the metric after a non-linear regression. This regression is performed using a 5-parameter logistic function as recommended in [14] and defined as follows:

$$Q_p(x) = \beta_1(\frac{1}{2} - \frac{1}{1+\exp^{\beta_2(x-\beta_3)}}) + \beta_4 x + \beta_5 \quad (4)$$

where β_i ($i \in \{1,2,3,4,5\}$) are five free-parameters to be fitted based on the Gauss-Newton method.

TABLE II
PERFORMANCE COMPARISON OF OBJECTIVE FIDELITY/QUALITY METRICS AND L_p NORMS DISTANCE METRICS.

Method	PLCC	SROCC	RMSE	OR
SSIM	0.936	0.939	0.416	0.152
MS-SSIM	0.858	0.942	0.677	0.152
VSI	0.876	0.955	0.634	0.138
VIF	0.925	0.932	0.500	0.172
VIFp	0.913	0.925	0.536	0.172
MAD	**0.977**	**0.973**	**0.275**	**0.119**
WSNR	0.941	0.936	0.445	0.158
FSIM	0.891	0.943	0.598	0.166
FSIMc	0.900	0.944	0.574	0.163
PSNR	0.962	0.958	0.357	0.138
UQI	0.901	0.907	0.571	0.186
IFC	0.922	0.914	0.509	0.166
NQM	0.942	0.936	0.441	0.172
GSIM	0.835	0.954	0.725	0.147
VSNR	0.923	0.917	0.507	0.186
L_0	0.885	0.915	0.613	0.186
L_2	0.914	0.958	0.533	0.138
L_∞	0.517	0.645	1.12	0.336

The PLCC, SROCC, RMSE and OR results are provided in Table II, where the top performing metric is given in boldface. Overall, a little more than half of the evaluated FR objective metrics provide good performance, especially MAD metric that shows the highest correlation with subjective scores.

As expected, the L_p distance metrics provide poor performance, except the L_2 distance that can be considered as acceptable, but still below those provided by the FR objective metrics. For instance, L_∞ distance has obtained the worst results compared to all evaluated metrics.

According to the reporting results, most of the FR objective metrics provide better performance than the widely used L_p distance metrics. Thanks to the inclusion of HVS features, the evaluated objective metrics correlate well with subjective scores and represent an obvious alternative to the L_p distance metrics. Consequently, the adoption and inclusion of FR objective metrics in the construction of adversarial attacks can produce more optimal results, thus allowing to contribute in developing more robust deep neural networks.

V. Conclusions

In this paper, we focused on the visual fidelity assessment of adversarial examples. We presented a publicly available dataset of adversarial examples, which can be used to the design and evaluation of new objective IFA metrics specifically developed for this kind of impairment. The dataset was constructed through subjective experiment, where the original images as well as adversarial examples, along with objective and subjective scores are provided.

The test results clearly exhibited that the L_p norms are non-suitable to quantify the perceived perturbations of adversarial examples, and that the objective fidelity/quality metrics represent a solid alternative to be a substitute for L_p norms.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Neural Information Processing Systems (NIPS)*, Nevada, USA, Dec. 2012, pp. 1097–1105.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[3] G. Hinton, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[4] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems (NIPS)*, pp. 3104–3112, Montral, Canada, Dec. 2014.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, San Diego, CA, May. 2015.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, Scottsdale, Arizona, USA, May. 2013.

[7] I. J. Goodfellow, N. Papernot, S. Huang, Y. Duan and P. Abbeel, "Attacking Machine Learning with Adversarial Examples," https://blog.openai.com/adversarial-example-research/. Open AI Blog. 2017.

[8] N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," *IEEE Access*, vol. 6, 14410–14430, Feb. 2018.

[9] X. Yuan, . He, Q. Zhu, X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[10] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (S&P)*, San Jose, CA, USA, May. 2017.

[11] Z. Wang and A.C. Bovik, "Mean squared error: love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[12] S. Winkler and P. Mohandas, "The evolution of video quality measurement: from PSNR to hybrid metrics," *IEEE Trans. Broadcast.*, vol .54, no. 3, pp. 660–668, Jun. 2008.

[13] S. Chikkerur, V. Sundaram, M. Reisslein and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol .25, no. 2, pp. 165–182, Jun. 2011.

[14] H. R. Sheikh, M. F. Sabir and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on image processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[15] D.A. Silverstein and J.E. Farrell, "The relationship between image fidelity and image quality," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Lausanne, Switzerland, Sep. 1996.

[16] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Workshop Track of the International Conference on Learning Representations (ICLR)*, Toulon, France, Apr. 2017.

[17] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu and J. Li, "Boosting Adversarial Attacks with Momentum," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, Jun. 2018.

[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in *International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 2018.

[20] N. Papernot *et al.*, "cleverhans v2.1.0: Adversarial Examples Library," arXiv preprint arXiv:1610.00768, 2018.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 2818–2826.

[22] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[23] ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," *International Telecommunication Union*, Apr. 2008.

[24] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," *Human vision and electronic imaging VIII*, Jun. 2003.

[25] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, Jan. 2012.

[26] ITU-R BT.2022, "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays," *International Telecommunication Union*, Aug. 2012.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[28] L. Zhang, L. Zhang, X. Mou, and D . Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[29] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[30] A. Liu, W. Lin, and M. Narwaria, "Image Quality Assessment Based on Gradient Similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, 2012.

[31] E. C. Larson and D. M. Chandler , "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, 2010.

[32] Z. Wang, E. P. Simoncelli, A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, 2003, pp. 1398–1402.

[33] L. Zhang, Y. Shen, and H. Li, "VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, Oct. 2014.

[34] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006

[35] H.R. Sheikh, A.C. Bovik and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[36] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002.

[37] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik, "Image quality assessment based on degradation model," *IEEE Trans. on Image Processing*, vol. 9, no. 4, pp.636–650, 2000.