Generalizability *vs.* Robustness: Investigating Medical Imaging Networks Using Adversarial Examples

Magdalini Paschali[1(✉)], Sailesh Conjeti[2], Fernando Navarro[1], and Nassir Navab[1,3]

[1] Computer Aided Medical Procedures, Technische Universität München, Munich, Germany
magda.paschali@tum.de
[2] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[3] Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, MD, USA

Abstract. In this paper, for the first time, we propose an evaluation method for deep learning models that assesses the performance of a model not only in an unseen test scenario, but also in extreme cases of noise, outliers and ambiguous input data. To this end, we utilize adversarial examples, images that fool machine learning models, while looking imperceptibly different from original data, as a measure to evaluate the robustness of a variety of medical imaging models. Through extensive experiments on skin lesion classification and whole brain segmentation with state-of-the-art networks such as Inception and UNet, we show that models that achieve comparable performance regarding generalizability may have significant variations in their perception of the underlying data manifold, leading to an extensive performance gap in their robustness.

1 Introduction

Deep learning is being increasingly adopted within the medical imaging community for a plethora of tasks including classification, segmentation, detection *etc.* The classic approach towards the assessment of any machine learning model revolves around the evaluation of its *generalizability i.e.* its performance on unseen test scenarios. However, in case of *limited* training data, such as medical imaging datasets, using heavily over-parameterized deep learning models could lead to the "memorization" of the training data. Evaluating such models on an available non-overlapping test set is popular, yet significantly limited in its ability to explore the model's resilience to outliers and noisy data/labels (*i.e.* robustness). Additionally, the limited interpretability of deep learning models due to their "black-box" nature challenges their adoption into clinical practice.

Existing model evaluation routines look deeply into over-fitting but insufficiently into scenarios of model sensitivity to variations of the input. Robustness

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11070, pp. 493–501, 2018.
https://doi.org/10.1007/978-3-030-00928-1_56

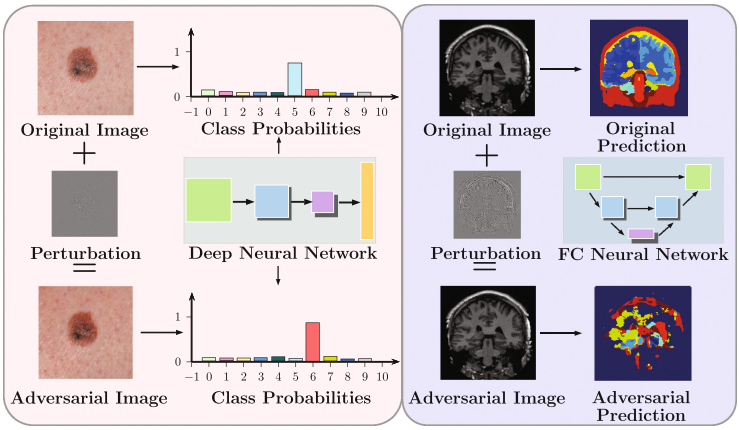


Fig. 1. Overview of Adversarial Crafting and its effect on network prediction. The difference between the generated adversarial image and the original image is imperceptible, yet deep neural networks are successfully fooled into anomalous predictions.

evaluation estimates potential failure probabilities when the model is *pushed* to its limits. In this paper, we approach evaluating a model by leveraging adversarial examples [1] that are crafted with the purpose of *fooling* a model and can uncover cases where its performance may degenerate. Our approach to using adversarial examples as benchmark is also significantly less laborious and expensive than constituting a sufficiently diverse test set with manual annotation.

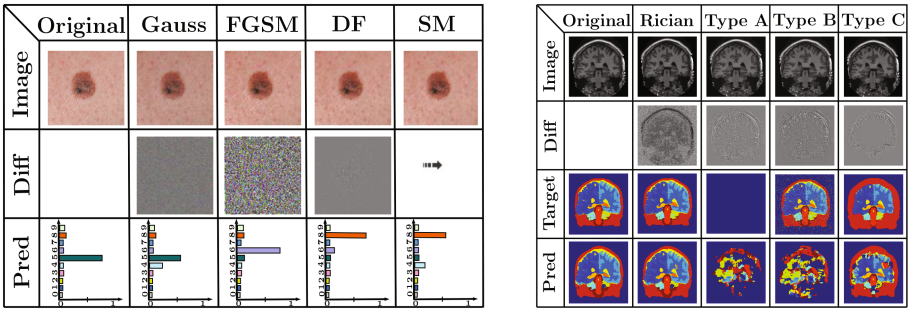
Adversarial examples are images crafted to purposely fool machine learning models, while the added perturbations are imperceptible to human eyes [1], as shown in Fig. 1. Our work is among the first that explore adversarial examples in medical imaging and leverage them in a constructive fashion to benchmark model performance not only on clean and noisy but also on adversarially crafted data. Previously, Zhu et al. augmented their dataset with adversarial examples to control overfitting and improve their model's performance on mass segmentation [2]. Even though adversarial examples may not occur in naturally acquired data, utilizing them can present new opportunities for medical imaging researchers to investigate their models, with the ultimate goal of increasing robustness and optimizing the decision boundaries learned for different tasks.

Our contribution is two-fold: Firstly, we demonstrate on a variety of medical image computing tasks that widely adopted state-of-the art deep learning models are not immune to adversarial examples crafting. Secondly, we utilize adversarial examples to benchmark model robustness by comparing a variety of architectures, such as Inception [3] and UNet [4], for the tasks of skin lesion classification and whole brain segmentation.

2 Methodology

2.1 Adversarial Crafting

Given a trained model F, an original input image X with output label Y, we generate an adversarial example $\hat{X}$ by solving a box-constrained optimization problem $\min_{\hat{X}} \|\hat{X} - X\|$ subject to $F(X) = Y$, $F(\hat{X}) = \hat{Y}$, $\hat{Y} \neq Y$ and $\hat{X} \in [0, 1]$. Such an optimization minimizes the added perturbation, say r (*i.e.* $\hat{X} = X + r$) while simultaneously *fooling* the model F [1]. By imposing an additional constraint such as $\|r\| \leq \epsilon$, we can restrict the perturbation to be small enough to be imperceptible to humans.



(Left) (Right)

Fig. 2. Illustration of adversarial examples and their effect on model predictions. Left: Skin lesion classification and Right: Whole brain segmentation. The arrow in the case of the SM attack indicates the few pixels that were perturbed. Contrasting with prediction on original images, the crafted examples are able to successfully *fool* the models into either misclassification or generating incorrect segmentation maps.

Classification: Gradient-based adversarial example generation methods have been proposed with the objective of generating minimum amount of perturbation r that misclassifies $\hat{X}$. These include the Fast Gradient Sign Method (FGSM) [5], DeepFool (DF) [6], Saliency Map Attacks (SMA) [7] *etc.* Adversarial examples crafted with these methods are shown in Fig. 2. For a trained model F, FGSM performs a one-step pixel-level update along the sign of the gradient that maximizes the task loss J and the resultant perturbation is computed as $r = \epsilon \text{sign}\left(\nabla_X J(\theta, X, Y)\right)$, where θ are the parameters of the model. The amount of perturbation is regulated by a hyper-parameter ϵ that is typically assigned a low value, so that $\hat{X}$ is visually imperceptible from X.

Differing from FGSM, DF follows an iterative greedy search process, where in each iteration the projections of the input sample to the decision boundaries of all the classes are computed and an r is calculated that will push X towards the closest decision boundary of a class, other than the correct one. In SMA, the

impact of each pixel on the prediction of the model is estimated and the input is selectively perturbed to cause the most significant change to the output.

Segmentation: In [8], the authors introduced Dense Adversarial Generation (DAG) as a method for crafting adversarial examples for semantic segmentation, closely resembling per-pixel, targeted FGSM. Particularly, DAG utilizes an incorrect segmentation mask, given by the user, and a target set of non-background pixels. Its goal is to calculate a minimum perturbation r that will alter the prediction from the correct class to the incorrect target class.

We utilized DAG to craft adversarial examples, seen in Fig. 2, by creating targets with varying degrees of difficulty. Particularly, we set the target to be all background (Type A), randomly assign a small percentage of pixels to a randomly-selected adversarial class (Type B) and modify (dilate) only a particular target class while keeping all other classes intact (Type C). Of the aforementioned attack types, Type A is the most challenging, causing the largest amount of perturbation, while Type C is expected to distort the image the least, as can be seen in Fig. 2. The Mean Square Error (MSE) between the original and adversarial images remained extremely small, ranging from 0.004 for adversaries of Type A to 0.002 for B and C.

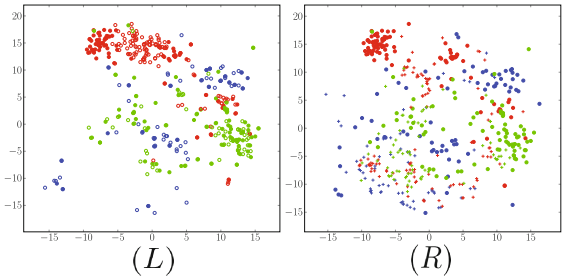


Fig. 3. t-SNE representation of the embeddings of 3 classes (red, blue and green) from clean (•), noisy (∘) and adversarial images (+). The noisy examples (∘) are embedded closer to clean data (L), while adversarial ones are *pushed* to the model boundaries (R).

2.2 Model Evaluation with Adversarial Examples

The proposed pipeline for the evaluation of robustness involves benchmarking models against task-specific adversarial attacks and is similar across tasks. For classification, we crafted adversarial examples with FGSM, DF and SMA, while for segmentation we applied DAG with 3 different types of targets (Type A−C). Afterwards we attacked our models in a black-box fashion with examples generated by independently trained models, to maintain an unbiased attack scenario.

Contrasting with Noise: One could argue that applying noise on the test images before inference could replace the need for adversarial examples. However,

that is not the case since hard ambiguous cases and outliers cannot be modeled by noise distributions. Adversarial examples, which are crafted with the purpose to force models to fail, are better suited for evaluating model behavior when subject to input extrema. To showcase that adversarial perturbations do not resemble noise distributions, we also crafted images distorted with modality-specific noise (Gaussian noise for dermatoscopic images and Rician noise for T1w MRI). For fairness, the Structural Similarity (SSIM) between the original and noisy images was the same as the one between the original and adversarial examples and ranged from 0.97 to 0.99.

We plot the t-Stochastic Neighbor Embedding representation (t-SNE) from IV3 for the clean, noisy and adversarial examples (FGSM) in Fig. 3 for the classification task to further illustrate this difference. Contrasting Fig. 3 (L) with Fig. 3 (R), we clearly observe that images distorted with noise are embedded close to the clean images, while adversarial examples are pushed further towards other classes. The anomalous nature of the adversarial examples clearly supports our hypothesis that their behavior is not akin to noise and can act as a harder benchmark for evaluating a model's robustness.

3 Experiments

To provide a proof-of-concept for the proposed robustness evaluation we chose the challenging tasks of fine-grained skin lesion classification and segmentation of the whole brain. The task-specific model learning is described as follows:

Classification: We fine-tune three state-of-the art deep learning architectures namely, InceptionV3 (IV3) [3], InceptionV4 (IV4) [3] and MobileNet (MN) [9] for this task. Both IV3 and IV4 are very-deep architectures (>100 layers), while MN is significantly compact. Comparing these architectures would help discover if any innate relationships exist between model complexity (in terms of depth and parameters) and their robustness. To keep the comparisons fair, all the models were initialized with their respective ImageNet parameters and fine-tuned with a weighted cross-entropy loss with affine data augmentation. Specifically, the models were trained with stochastic gradient descent with a decaying learning rate initialized at 0.01, momentum of 0.9 and dropout of 0.8 for regularization. We use the publicly-available Dermofit [10] image library consisting of 1300 high-quality dermatoscopic images, with histologically validated fine grained expert annotations (10 classes) for this task. The dataset was split at patient-level with non-overlapping folds (50% for training and rest for testing).

Segmentation: For this task we chose to evaluate three popular fully-convolutional deep architectures, namely SegNet (SN) [11], UNet (UN) [4] and DenseNet (DN) [12]. Contrasting across these architectures, we evaluate the importance of skip connections with respect to robustness varying from no skip connections in SN to introducing long-range skips in UN and both long and short-range skip connections in DN. The model parameters (depth and layers) were chosen to maintain comparable model complexity, so as to exclusively factor out

Table 1. Comparative evaluation of the classification and segmentation models on clean, noisy and adversarial examples. We report the average accuracy and Dice overlap score along with the % drop in performance on adversarial examples with respect to performance on clean data.

			Noise	Adversarial	
Classification		**Clean**	**Gaussian**	**Avg**	**% Drop**
	IV3 [3]	0.710	0.693	**0.641**	**6.897**
	IV4 [3]	**0.810**	**0.761**	0.633	17.72
	MN [9]	0.800	0.647	0.564	24.55
Segmentation		**Clean**	**Rician**	**Avg**	**% Drop**
	SN [11]	0.842	0.595	0.470	37.17
	UN [4]	**0.862**	0.759	0.453	40.92
	DN [12]	0.861	**0.848**	**0.667**	**19.53**

the impact of skip connections to robustness. The aforementioned models were trained with a composite loss of weighted-cross entropy and Dice loss [13] and model optimization was performed with ADAM optimizer with an initial learning rate of 0.001. We use 27 volumes from the publicly-available whole-brain segmentation benchmark (subset of Open Access Series of Imaging Studies (OASIS) dataset [14]) that was released as a part of the Multi-Atlas Labeling Challenge in MICCAI 2012 [15], with 80-20 patient-level splits for training and testing. The models for both tasks were trained until convergence using TensorFlow [16] and adversarial examples for DF and SMA attacks described in Sec. 2.2 were crafted using the FoolBox [17] library.

Following the model evaluation strategy presented in Sec. 2.2, adversarial examples were crafted for each of the trained models and their robustness is evaluated in terms of average classification accuracy and average Dice score. We report the overall performance of the models in Table 1, where we compare the performance of each model on clean and noisy images with their average score against all the attacks. Furthermore, in Table 2 we are reporting the performance of each model against all the black-box attacks separately.

4 Results and Discussion

4.1 Robustness Evaluation for Classification

Visual Evaluation: Fig. 2 (Left) illustrates adversarial examples crafted for an unseen test example (belonging to malignant melanoma class) for each of the classification related attacks (FGSM, DF and SMA) alongside an image perturbed with Gaussian noise for comparison. A scaled version of the difference image with respect to the original is also shown along with the posterior probabilities estimated using IV3 network. We observe that all the adversarial examples, regardless of the attack are consistently misclassified with very high confidence,

while the addition of Gaussian noise only results in confidence reduction. Furthermore, FGSM induces perturbations dispersed across the whole image, while DF and SMA generate perturbations more localized to the lesion.

Attacks: From Table 1, we observe that IV4 and MN both achieve comparable performance on clean data (80–81%) superior to the one of IV3 (71%). By limiting model evaluation to generalizability (*i.e.* performance on clean data) one may prematurely conclude that IV3 demonstrates the worst comparative performance. However, upon comparing the robustness of these models with respect to average performance under all the attacks (Table 2), we observe a contrary trend. The performance drop for IV3 is significantly lower (7%) in comparison to IV4 (17%) and MN (25%). IV4 achieves higher accuracy on DF and SMA attacks, while IV3 is the most robust model against FGSM. Contrasting IV4 and MN, we observe that MN performs poorly not only on noisy samples but also on all of the attacks, as shown in Table 2. These contrasting observations clearly substantiate the core hypothesis within the paper that model evaluation should not be limited to generalizability and that performing robustness evaluation is equally important. Despite showing comparable performance in terms of generalizability, we can clearly conclude that IV4 is strongly preferred over MN.

Table 2. Comparative evaluation of model robustness using black-box attacks for the tasks of classification and segmentation. We report the average accuracy for classification and average Dice overlap score across structures for segmentation.

		FGSM			DF			SMA		
		IV3	IV4	**MN**	IV3	IV4	MN	IV3	IV4	MN
Classification	**IV3** [3]	**0.449**	**0.548**	**0.567**	0.729	0.707	0.664	**0.738**	0.701	0.669
	IV4 [3]	0.429	0.411	0.451	**0.743**	**0.768**	**0.697**	0.735	**0.778**	**0.683**
	MN [9]	0.335	0.275	0.213	0.726	0.731	0.672	0.732	0.735	0.661
		Type A			Type B			Type C		
		SN	UN	DN	SN	UN	DN	SN	UN	DN
Segmentation	**SN** [11]	0.277	0.272	0.309	0.397	0.473	0.428	0.669	0.702	0.705
	UN [4]	0.248	0.434	0.258	0.364	0.434	0.368	0.636	0.653	0.677
	DN [12]	**0.600**	**0.528**	**0.415**	**0.749**	**0.721**	**0.563**	**0.819**	**0.791**	**0.814**

4.2 Robustness Evaluation for Segmentation

Visual Evaluation: Fig. 2 (Right) illustrates how the prediction maps of the trained DN model transform when it is attacked. All the DAG attacks (Type A-C) successfully fool the model into producing an incorrect prediction map. However, the prediction on the image distorted with Rician noise is visually similar

to the one of the original image and the ground truth. This clearly demonstrates that adding adversarial perturbation is not akin to adding random noise.

Attacks: From Table 1, we observe that DN (86.1%) and UN (86.2%) achieve almost identical performance on clean unseen test examples and fare better than SN (84.2%), highlighting the importance of skip connections. Furthermore, the fact that the performance drop caused by the addition of Rician noise remains low for UN and DN (10% and 1% respectively) reinforces the distinction between noise and adversarial perturbations. Regarding model performance with respect to adversarial attacks in Table 2, we observe that DN is not only resilient to noise but also significantly more robust than SN (by 18%) and UN (by 21%) to attacks crafted by any other model. Furthermore, SN and UN remain highly vulnerable to all the attacks with a significant 37–40% drop in their average Dice score. In consensus to the classification results discussed earlier, comparing the three segmentation models only in terms of generalizability would not have been sufficient to determine the best one. Both its resilience to samples distorted with Rician noise and its consistent resistance to adversarial attacks make DN the strongest model among its competitors for this task.

5 Conclusion

In this paper we explored adversarial examples in medical imaging for the tasks of classification and segmentation and proposed a strategy for model evaluation by leveraging task-specific adversarial attacks. We showed that for two models with comparable performance, their relative exploration of the underlying data manifold may have significant differences, hence resulting in varying robustness and model sensitivities. Specifically, we demonstrate that for segmentation tasks the use of dense blocks and skip connections contributes to both improved generalizability and robustness, while model depth seems to increase the resistance of classification models to adversarial examples.

References

1. Szegedy, C., et al.: Intriguing properties of neural networks. In: ICLR (2014)
2. Zhu, W., Xiang, X., Tran, T.D., Hager, G.D., Xie, X.: Adversarial deep structured nets for mass segmentation from mammograms. In: ISBI (2018)
3. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
4. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: MICCAI (2015)
5. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
6. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: CVPR (2016)
7. Papernot, N., McDaniel, P.D., Jha, S., Fredrikson, M., Berkay Celik, Z., Swami, A.: The limitations of deep learning in adversarial settings. In: EuroS&P (2016)

8. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.L.: Adversarial examples for semantic segmentation and object detection. In: ICCV (2017)
9. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. CoRR abs/1704.04861 (2017)
10. Ballerini, L., Fisher, R.B., Aldridge, R.B., Rees, J.: A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Color Medical Image Analysis (2013)
11. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
12. Jégou, S., Drozdzal, M., Vázquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: fully convolutional DenseNets for semantic segmentation. CVPR Workshops (2017)
13. Roy, A.G., Conjeti, S., Sheet, D., Katouzian, A., Navab, N., Wachinger, C.: Error corrective boosting for learning fully convolutional networks with limited data. MICCAI (2017)
14. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosc. **19**(9), 1498–1507 (2007)
15. Landman, B., Warfield, S.: MICCAI workshop on Multiatlas labeling. In: MICCAI Grand Challenge (2012)
16. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al.: TensorFlow: large-scale machine learning on heterogeneous distributed systems. CoRR abs/1603.04467 (2016)
17. Rauber, J., Brendel, W., Bethge, M.: Foolbox v0.8.0: A Python toolbox to benchmark the robustness of machine learning models. CoRR abs/1707.04131 (2017)