# DRAFT: Robustness in Medical AI: A training scheme for more robust semantic segmentation

Konstantin Avramidis[a,*], Peter Kieseberg[b], N.N.[a], Andreas Holzinger[a]

[a] *Medical University Graz, Austria*
[b] *Institute of IT Security Research, St. Pölten, Austria*

**Abstract**

Convolutional Neural Network Classifiers are well studied and recent advancements in hardware technology made them applicable to many day to day tasks and therefore a popular target for malicious attacks. These attacks, called *adversarial attacks*, compared to the inception of such Networks, are fairly new and an ongoing field of research. Many of the approaches to fortify predictions against such attacks rely on the obfuscation of gradient surfaces and only provide some protection which in turn can be overcome by applying more computing power or smarter algorithms. This hints at some inherent design flaw in the architecture of Convolutional Neural Networks which may be not entirely protectable. **However, epistemology taught us that error produces a deeper understanding of the matter.** In some sense this is mimicked by Generative Adversarial Networks which try to reach an equilibrium of both failure reduction and induction. In order to be truly intelligent an agent needs to be able to formulate an hypotheses and an experiment to either falsify or confirm such hypotheses. Therefore this work concentrates on setting up an environment where a semantic segmentation agent is able to falsify or confirm its prediction by applying a hypotheses, at least to some extend, in form of an adversarial attack to the data given to it.

*Keywords:* Explainable AI, explainability, interpretable Machine Learning,

*Corresponding author
*Email address:* andreas.holzinger@medunigraz.at (Andreas Holzinger)

Glass-box approach, transparency, interpretability

**Article Summary**

Inspired by the work of [1, 2, 3] and [4] we try to further analyse findings presented by mentioned works. This analysis will not be conducted on CIFAR or ImageNet data sets, but is carried out in the context of image segmentation. Hereon custom U-Nets [5] are used with an openly available image set of histological data from Kaggle which seemed fit for the hardware available to us which is a single machine utilizing one Geforce 1080 ti. In this setup we produce different perturbations by applying different schemes of Fast Gradient Sign Methode (FGSM) [4] to training sets to increase robustness, as well as accuracy. We are fully aware of the work of [6], as well as [7]. To some extend this work will extend these papers by presenting alternative training schemes which not only use adversarial perturbation, but also feature enhancing perturbation to strengthen segmentation accuracy. **Our intent is by presenting a network both kinds of perturbations a selection process is triggered that filters brittle features and sustains robust ones.**

The contribution of this work lies in:

- 1) Providing a Python Library that supplies an easy to use segmentation algorithm
- 2) Executing various attacks against said algorithm
- 3) Providing suggestions on how to use said attacks against the items of the provided Library
- 4) Analyzing improvements of robustness and accuracy by applying said attacks against the selected algorithm.

## 1. Introduction

The formalization of adversarial examples [8, 9] in Deep Neural Networks (DNN) has led to a manifold of problems. One of the biggest is that a commer-

cially viable product needs to be save to use in the intended field of application. While the international robustness research community is mainly concentrating on image classification tasks, the subject of Image Segmentation does get very little attention [10]. This seems to stem at least to some extend from the fact that the Generative Adversarial Nets (GAN) [11] community is focusing more on the topic and GANs are in a lot of points similar to adversarial attacks, therefore leaving little room for security research to examine their concerns since adversarial behavior is deemed to be features in the context of GANs. Furthermore, the public availability and feasibility of high quality data sets is very limited. Mainly the cityscapes [12] data set is used, which provides 5000 high resolution images of inner-city traffic scenery and corresponding fine annotated masks of 30 classes. Since analysis of such a data set requires considerable compute power for day to day research, it is not as low entry as MNIST [13] or even CIFAR [14] classification. While contemporary critical commercial applications mainly rely on standard classification, future prospect suggests the wide-spread use of image segmentation, as well as GANs in security sensitive environments like self driving cars or medical diagnostics. Therefore, it is most imperative to test and develop robust frameworks to provide as secure image segmentation tasks as possible. The FGSM-attack concept is a whitebox attack that is utilizing the networks parameters to compute a stochastic step, which increases the loss function value with respect to the target. By solely evaluating the signum-function of this step, it is possible to get a directional matrix as shown in Formula 1 within the input-space, which when added to the input perturbs it in a way that causes the classification to push predictions into another class. There are derivatives of this attack like Projected Gradient Descent (PGD) [2]. PGD is basically an iterative procedure using FGSM in a multi-step setting, accumulating steps onto the input. Interestingly enough, both methods can be used to do the opposite, meaning it is possible to fortify predictions by adding perturbation through simply changing an addition to a subtraction or declaring an artificial target. In this context, this paper uses said attacks to create a data set that not just includes perturbed input examples, but also fortified ones to

further strengthen the ability of neural networks to choose robust features in training by hypotheses generated in a previous iteration of the network about what resemblance a class and what can be viewed as static or noise. We leave it to future work to integrate steps into conventional training that would mimic such data augmentations to reduce training time as well as training cycles.

## 2. Methods

To generate different qualities of adversarial examples we use two methods in this work: First we produce a layered iterative adversarial example on bases of PGD taking 100 steps that enforces full area segmentation classes in every iteration, so that the desired target mask represents a single class. This is continued as long as there are different classes. In every pass, the next iteration is calculated on bases of the previous example as described in Formula 2. **This approach ensures that non distinctive patterns are encoded within the adversarial image. This is illustrated in Figure 1 which shows a fragmented prediction caused by the patterns encoded by a previous step. If PGD would be applied in a traditional fashion the attack would produce a full area prediction like in Figure 2. Adding this examples into the trainings set a network should be able to recognize these over expressed patterns as unhelpful, which in turn will give a better and more robust segmenting network.**

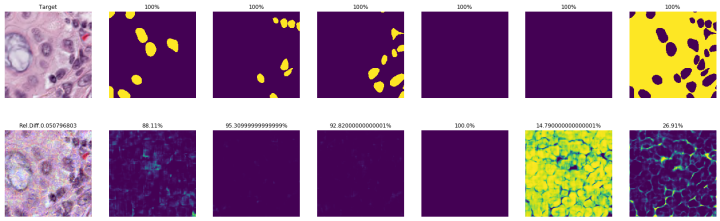


Figure 1: Sample of transformation and effect achieved by method one vs. training example and target

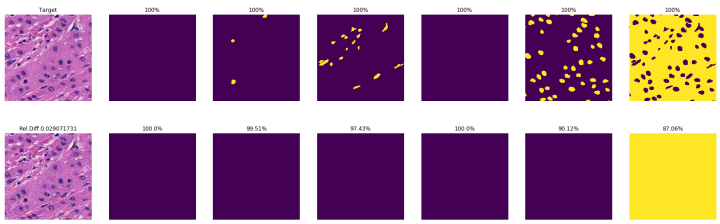

Figure 2: Sample of transformation and effect achieved PGD vs. training example and target

$$\eta_{x,y} = \varepsilon * sign(\nabla_x J(\Theta, x, y)) \tag{1}$$

$$X_{adv+1} = clip(X_{adv} + \sum_{adv=1}^{||Y_{adv}||} -\eta_{X_{adv},Y_{adv}} \tag{2}$$

Secondly, we deploy an algorithm that selectively produces enhanced examples on basis of their overall prediction score. It is allowed to produce FGSM perturbations as described by Formula 1 but with the freedom of choosing if they should be added, removed or not be considered at all for the example. This is achieved by alternating the signum inside the formula and then letting the network's loss function 3 choose which combination of perturbations will enhance predictions. Thereby, non-expressive class-specific patterns are removed from the data while strongly classifying ones are enhanced as expressed by Formula 5.

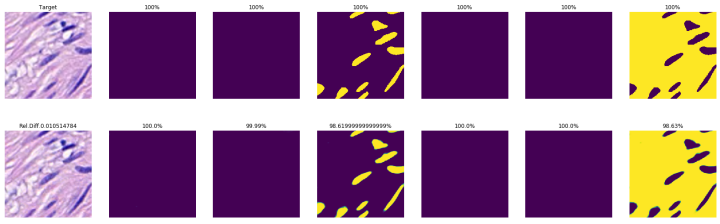

Figure 3: Sample of transformation and effect achieved by method two vs. training example and target

$$l_{x,y} = \nabla_x J(\Theta, x, y) \tag{3}$$

$$a = \eta_{x,y_{adv}} \tag{4}$$

$$x_{imp} = \begin{cases} x + a & \text{if } min(l_{x+a,y}) < l_{x,y} \\ x & \text{if } l_{x,y} <= min(l_{x\pm a,y}) \\ x - a & \text{if } l_{x,y} > min(l_{x-a,y}) \end{cases} \tag{5}$$

Both concepts are very time-consuming in their current implementation. The first one scales linearly and is heavily depended on the amount of iterations PGD takes. The second behaves exponentially to the bases of 3. That means in the case of six classes there are 3 to the power of 6 combinations that need to be tested for every image in the training set. These 729 permutations need to be generated by and evaluated against the U-Net, therefore not only the permutational complexity, but also the extend of the underlying network needs to be taken into account, even though in the case of the second method the complexity of a single step FGSM seems negligible compared to the 100 steps of PGD. Both resulting data sets resemble the original data set. To further test the behavior and confirmation of enhancement, two separate networks are trained, each on it's own data set. They are evaluated against the standard test set and show similar accuracy as well as predictions. This confirms the ability of the U-Net to see passed the adversarial signal. Here we need to undertake more thorough research to confirm how exactly the two different approaches impact the behaviour of semantic segmentation algorithms. In a last step we generate a fourth network on bases of the previously calculated perturbations, as well as the original training set. To accommodate for the growth of training data we reduce the epochs by two thirds.

## 3. Experiments

As mentioned before, four networks were created: Model one is a costum U-Net trained with the previously mentioned histological data set. Model two will be trained with data perturbed by method one. Model three will be trained by data augmented by method two. And finally model 4 will be trained by a combination of the three. Their Outputs are evaluated against PGD with 100 steps targeting full area background class and a negative epsilon of 1‰. This is a very powerful attack as confirmed by Figure 2 and 5. Clearly expressed and fully visible nuclei are not recognized by the network. It should be noticed that the base reading of the unenhanced model one against the previously explained PGD attack is an accuracy of around 80%. This is easily explained by the fact that the pixels of the 5 semantic classes together hold only 3.2% of the values while the background class pixels hold around 81.6%. Choosing the background class as comparative value still makes sense under the assumption of an attack. If there is nothing to see then interest will be low.

Table 1 shows comparable values for all four model's accuracy when tested against the unperturbed data set. Even though the different loss values seem to be of significance, all four models show fluctuation during training as Figure 4 documents. It is not entirely clear why this (besides obviously overfitting) happens but the author speculates that hyperparameter tuning in the sense of an adaptive learning rate could be helpful. We also observed imprecision associated with the data labeling which could be causing said fluctuations.

| U-Net | loss | accuracy |
|---|---|---|
| 1 | 0.4959 | 0.8748 |
| 2 | 0.5029 | 0.8655 |
| 3 | 0.8410 | 0.8769 |
| 4 | 0.7059 | 0.8738 |

Table 1: Test Results on unperturbed and unseen 561 Images

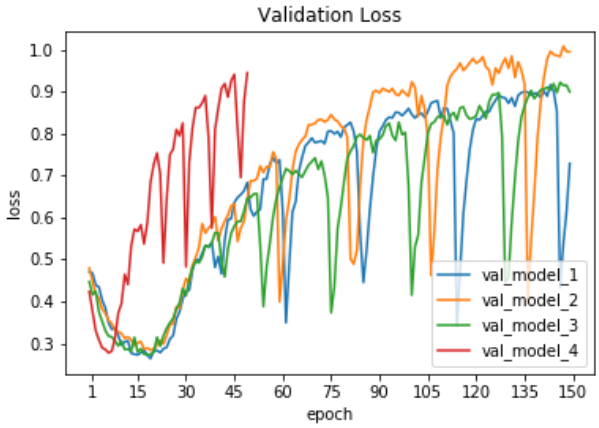


Figure 4: Graph validation loss for all four Models

In Table 2 we observe a basically unchanged accuracy for model four shown in Figure 8. Although the complete test set comprises of perturbed images, model four even surpasses model two, which was trained only on adversarial examples, by a slight margin. This seems counter-intuitive at first glance, but makes sense when considering how method two operates. Although not trained on the PGD perturbations, the network seemingly managed to filter for meaningful patterns explaining the not perfect but, considering the limitations, impressive result shown in Figure 7 representing model three. Particularly in direct comparison with Figure 6 showing the output of model two which was trained only on PGD perturbations.

| U-Net | loss | accuracy |
|---|---|---|
| 1 | 2.3628 | 0.8087 |
| 2 | 0.5068 | 0.8662 |
| 3 | 1.4774 | 0.8386 |
| 4 | 0.7205 | 0.8726 |

Table 2: Test Results on perturbed and unseen 561 Images

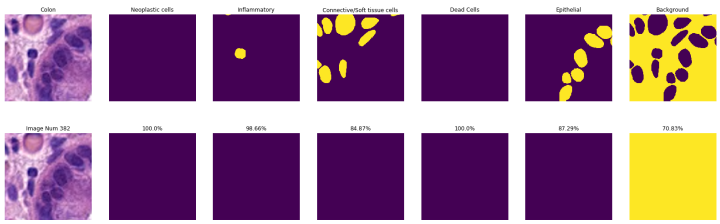


Figure 5: Prediction of Model 1 against a 100 Step PGD with Epsilon of 1‰

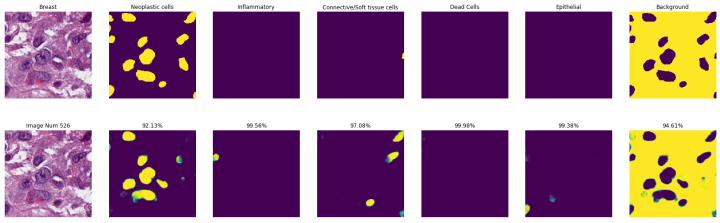


Figure 6: Prediction of Model 2 against a 100 Step PGD with Epsilon of 1‰

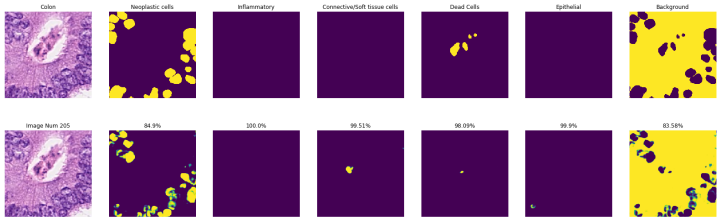


Figure 7: Prediction of Model 3 against a 100 Step PGD with Epsilon of 1‰

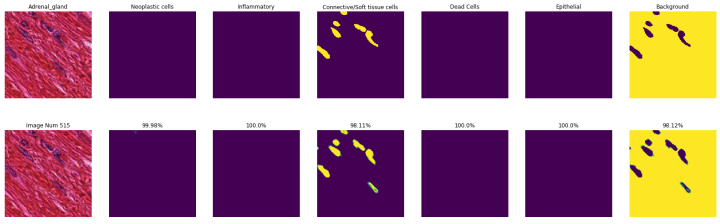


Figure 8: Prediction of Model 4 against a 100 Step PGD with Epsilon of 1‰

## 4. Discussion and Conclusion

Although network four presents robustness and a little better accuracy, interpretation of results is hard. For one, a metric to measure true enhancement is hard to come by since a picture natively is comprised of at least one class and the background class. Shifts in segmentation that include more classes can be positive in terms of accuracy, because they reduce error in the background class, but can also lead to misclassification in terms of semantic classification. This paper used the mean error to compensate for discrepancy over all classes to determine accuracy in method two. There maybe better ways to account for mutliclass-errors, but due to time- as well as computational-limitations we used the mean point for point accuracy. Another problem is the quality of the used data. High quality semantic segmentation data is not readily available yet and during testing semantic coherence seemed to be not completely trustworthy. This was made evident by viewing training graphs observing fluctuations of validation loss. That is why we recommend Kandinsky Patterns [15] for subsequent research. Nonetheless, we believe that method two improved behaviour of adversarial training and is worth a more rigorous investigation.

**The code used to generate and corroborate results presented here, misses an option to document choices made by method two in regards of which changes were made to the input data for augmentation.** In our opinion, this is an interesting topic for further investigation since they should have a viable impact on decision-boundaries of adversarially trained descendants of the original network as indicated by model 3, which showed partial robustness against PGD.

Furthermore we like to emphasise that the inherent brokenness of all CNNs will not be solved by adding perturbations [16] into training routines, transforming inputs [17] until perturbations lose their effectiveness nor obfuscating gradients [18] in what ever form. Especially in fields where the utmost precision is required like medicine or decision dependant automation. Therefore more research or even a completely new approach for neural network design is

needed.

**Acknowledgements**

We are grateful for interesting discussions with our local and international colleagues and their encouragement. This research was funded by the European Union's Horizon 2020 research and innovation program under grant agreement No.826078, the Austrian Science Fund (FWF) through project P-32554 "explainable Artificial Intelligence" and the Austrian Research Promotion Agency (FFG) trough the COIN project 866880 "Big Data Analytics (BDA)".

## References

[1] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, A. Madry, Adversarial examples are not bugs, they are features, in: Advances in Neural Information Processing Systems, 2019, pp. 125–136.

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083.

[3] J. Hendrik Metzen, M. Chaithanya Kumar, T. Brox, V. Fischer, Universal adversarial perturbations against semantic image segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2755–2764.

[4] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, stat 1050 (2015) 20.

[5] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[6] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, Robustness may be at odds with accuracy, arXiv preprint arXiv:1805.12152.

[7] P. Nakkiran, Adversarial robustness may be at odds with simplicity, arXiv preprint arXiv:1901.00532.

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199.

[9] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, F. Roli, Evasion attacks against machine learning at test time, in: Joint European conference on machine learning and knowledge discovery in databases, Springer, 2013, pp. 387–402.

[10] C. Kamann, C. Rother, Benchmarking the robustness of semantic segmentation models, ArXiv abs/1908.05005.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, CoRR abs/1604.01685. arXiv:1604.01685.
URL http://arxiv.org/abs/1604.01685

[13] Y. LeCun, C. Cortes, C. Burges, Mnist handwritten digit database.

[14] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images.

[15] A. Holzinger, M. Kickmeier-Rust, H. Müller, Kandinsky patterns as iq-test for machine learning, in: International cross-domain conference for machine learning and knowledge extraction, Springer, 2019, pp. 1–14.

[16] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, T. Goldstein, Adversarial training for free!, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 3358–3369.
URL http://papers.nips.cc/paper/8597-adversarial-training-for-free.pdf

[17] C. Guo, M. Rana, M. Cisse, L. Van Der Maaten, Countering adversarial images using input transformations, arXiv preprint arXiv:1711.00117.

[18] A. Athalye, N. Carlini, D. Wagner, Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, arXiv preprint arXiv:1802.00420.