Towards Robustness in Medical AI: A training scheme for more secure semantic segmentation

Konstantin Avramidis, Peter Kieseberg, Rudolf Mayer, Edgar Weippl, and Andreas Holzinger,

Abstract—Deep Learning classifiers are well studied and are applied to many tasks, and therefore a popular target for malicious attacks. Adversarial attacks, compared to the inception of such Networks, are recent, 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. To be truly intelligent, an agent needs to be able to formulate a hypotheses, and an experiment to either falsify or confirm such hypotheses. This work concentrates on setting up an environment where a semantic segmentation agent is able to falsify or confirm it's prediction by applying a hypotheses, in form of an adversarial attack to the data given to it, in contrast to it's prediction made after a first training cycle.

Index Terms—Robustness, Medical AI, Machine Learning

1 Introduction

The formalization of adversarial examples [1], [2] in Deep Neural Networks (DNN) has led to a manifold of problems. One of the biggest issues that arose from this work, is that a commercially viable AI product needs to be safe to use and trust in the intended field of application [3], like self driving cars or high-level automation in medicine, even in the presents of such attacks. Adversarial examples pose a threat in this regard because they can trigger unwanted behaviour with, for humans, unrecognizable perturbations. While the international research community on adversarial machine learning is mainly concentrating on image classification tasks, the subject of Image Segmentation until now received very little attention [4]. This seems to stem, at least to some extent, from the fact that the Generative Adversarial Nets (GAN) [5] community is focusing more on the topic of inflicting adversarial behaviour, and GANs are in many ways similiar to adversarial attacks [6] as they are stochastic feature distillation processes. This leaves less room for security research to examine concerns since adversarial behavior related to GANs is deemed to be a feature. Therefor results presented in this context give the impression of success more easily as they only need to be inflicted but not countered. Another impediment to advance research on robustness in image segmentation is the public availability and obtainability of high quality data sets, which is very limited. Mainly the cityscapes [7] data set is used, which provides 5,000 high resolution images of inner-city traffic scenery, and corresponding fine annotated masks for 30 classes. Since analysis of such a data set requires considerable computing power, for day to day research it is not as suitable as MNIST [8] or even CIFAR [9] is for the image classification commuinity.

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 – though GANs will not play a big role in decision support/making processes, but rather synthetic feature generation for training routines [10]. Therefore, it is imperative to test and develop robust frameworks and deepen the understanding of the matter to provide as secure as possible image segmentation solutions.

The FGSM-attack (Fast Gradient Sign Method) [11] 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 sign-function of this step, it is possible to obtain a directional matrix as shown in Equation (1) within the input-space, which, when added to the input, perturbs it in a way that causes the classification to push predictions over the decision boundary, into another class. There are other derivatives of this attack, such as the Projected Gradient Descent (PGD) [12]. 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 and thereby reducing loss, or declaring an artificial target so that the resulting directional matrix should step away from a position were ϵ can be attacked by a small perturbation. In this context, this paper uses said attacks to create a data set that not just includes perturbed input examples [12], but also fortified ones. This is to further strengthen the ability of neural networks to choose robust features in training by hypotheses generated

Corresponding author: Andreas Holzinger.

- *Andreas Holzinger is with the Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Canada and head of the Human-Centered AI Lab, Medical University Graz, Austria. E-mail: andreas.holzinger@medunigraz.at.*

in a previous iteration of the network about what resemblance a class of a single pixel in the context of semantic segmentation and what can be viewed as static or noise. We would be happy if fellow researches felt encourage by this work to integrate steps into conventional training that would mimic such data augmentations to reduce training time as well as training cycles now necessary to accomplish described setup.

1.1 Our Contributions

Inspired by the work of [12], [13], [14] and [11] concerning the nature and properties of adversarial examples we seek to further analyse findings presented by mentioned publications. This analysis will not be conducted on CIFAR or ImageNet data sets, but instead will be carried out in the context of image segmentation. Hereon custom U-Nets [15] are used, which, despite what the scheme of Ronnenberger et al. [15] suggests, use padding. These neural nets are trained with an openly available image set of histology data for both segmentation and classification tasks, introduced by [16] and made available e.g. on Kaggle[1], which seemed fit for the hardware available to us. Thereby experiments are designed to be carried out on a single machine makeing them highly reproducible. In this setup, we create different perturbations by applying different schemes of Fast Gradient Sign Method (FGSM) to training sets to increase robustness, as well as accuracy. We are fully aware of the work of [17], as well as [18] and [19]. Their work implies that there are no completely protectable configuration for highly accurate neural networks. To some extent, this work will extend these papers by presenting alternative training schemes, which not only use adversarial perturbation, but also feature enhancing perturbation that reduces loss for the target mask y to strengthen segmentation accuracy. **This approach intents to create a learning environment, which, by presenting a network both perturbations and enhancements, triggers a selection process 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 and scale segmentation algorithm in form of a custom U-Net
- 2) Executing various attacks against said algorithm also provided by the code library
- 3) Providing suggestions on how to use said attacks against the items of the chosen data distribution.
- 4) Analyzing improvements of robustness and accuracy by applying said attacks against the selected algorithm.

1.2 Paper Organization

Section 2 provides a detailed look on how the two different altering processes, we use to trigger a feature selection process, demonstrated in Section 3, are designed, and how they are going to be utilized to achieve more robust daughter networks derived from the primary network with similar accuracies in both training as well as test error. Section 3 details experimentally validated results and behaviour achieved using said algorithms. Section 4 is going to put our work into context with previous and similar research. Finally, Section 5 will adhere to the fact that the data used over-proportionally comprises of background which gives a high accuracy when attacks that are peculiar to the background class being used.

1. https://www.kaggle.com/andrewmvd/cancer-inst-segmentation-and-classification

2 Methods

To generate different qualities of adversarial examples, we use two methods in this work. In the first method, we produce a layered iterative adversarial example, based on the Projected Gradient Descent (PGD) method taking 100 steps, that consistently enforces full area segmentation classes in every iteration for every available class in Y_{adv}, so that the desired target mask $\hat{y}$ represents a single class in every iteration. In Figures 3 and 8 an attack like this is shown for the background class. This is continued as long as there are different classes. In every pass, the next iteration is calculated on basis of the previous example, as described in Equation (2). This approach ensures that non-semantically distinctive patterns are encoded within the adversarial image because segmentation in the very meaning of the word is suppressed to approximate a classification problem, albeit an attacked one. This is illustrated in Figure 2, which shows a nearly full area class prediction for a sample seen before by the network. Fragments were caused by the patterns encoded by the previous 5 steps including background semantic class. If PGD would be applied in a traditional fashion, the attack would produce a full area prediction like in Figures 3 and 8, resembling the deterministic nature of standard classification, which only evaluates class affiliation. By adding these examples into the training set, a neural network should be able to recognize these over-expressed patterns as unhelpful, which in turn will yield a network more robust for the segmentation task. An unperturbed prediction from an distribution unknown to the network is shown in Figure 1 which indicates good generalization abilities for the five semantic classes of cell-nuclei types present in the distribution.

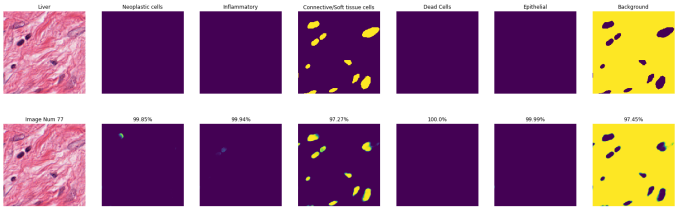

Fig. 1. Sample of prediction (lower row) achieved by model one against a sample of the test set (upper row).

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

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

Secondly, we developed an algorithm that selectively produces enhanced examples on basis of their overall prediction score. It is allowed to produce FGSM perturbations as

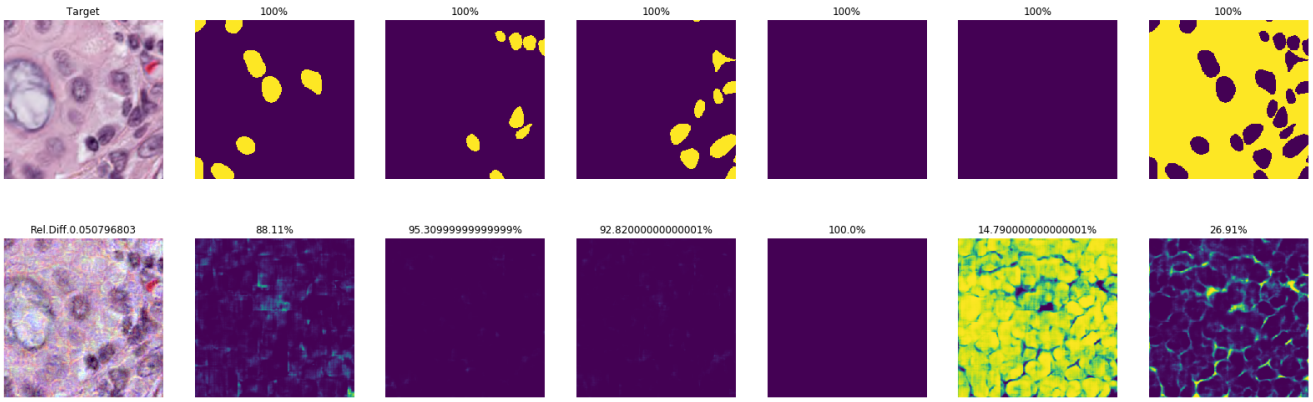

Fig. 2. Sample of transformation and effect achieved by method one (lower row) vs. model one training example and target (upper row).

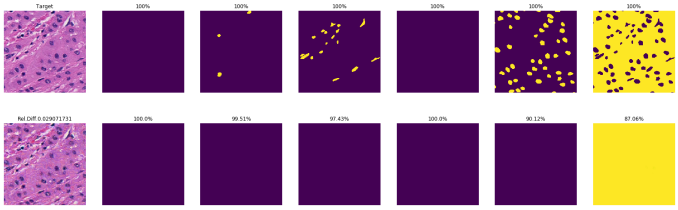

Fig. 3. Sample of prediction done by model one perturbed by 100 step PGD (lower row) against a sample of the training set (upper row).

described by Equation (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 equation by a multiplication with -1, and then letting the network's loss function Equation (3) choose which combination of perturbations will enhance predictions. Thereby, non-expressive but class-specific patterns are removed from the data, while strongly discriminating ones are enhanced as expressed by Equation (5).

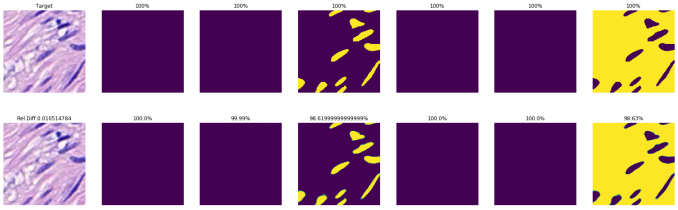

Fig. 4. Sample of transformation and effect achieved by method two (lower row) vs. model one training example and target (upper row).

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

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

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

Both methods are very time-consuming in their current implementation. The first one scales linearly to the number of classes, and is heavily dependent on the amount of iterations PGD takes. The second behaves exponentially to the base of 3 ($O(3^n)$). That means in the case of six classes ($n = 6$), including background class, 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 extent 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.

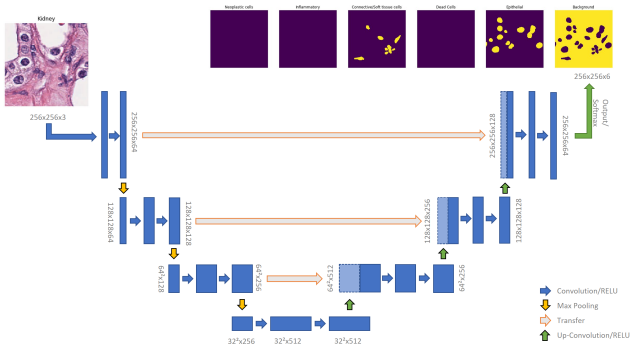

Fig. 5. Schema of U-Net Architecture as utilized in this work.

Both resulting data sets resemble the original data set. To further test the behavior and confirmation of our enhancement, two separate networks are adversarilly trained, each on one of the previously described augmented data sets. They are evaluated against the standard test set, and show similar accuracy as well as predictions $\hat{y}$. This confirms the ability of the U-Net to see past the adversarial signal after adversarial training. Here we need to undertake more thorough research to confirm how exactly the two different approaches impact the behaviour of semantic segmentation algorithms.This could be done by evaluating the absolute difference of the resulting vectors in contrast to the prediction of a non-adversarially trained network.

In a last step we generate a fourth network on the basis of both of the previously calculated enhanced data sets, as well

as the original training set. To accommodate for the growth of training data, we reduce the epochs by two thirds.

3 Experiments and Results

The intend is to create four neural networks which are to be trained by four different variation of the same data set. These variations are achieved by the two methods described in Section 2. All four models are derived from a custom U-Net[2] architecture schematically depicted in Figure 5. It is different, as indicated before, from the standard scheme proposed by Ronneberger et al. by using padding to prevent edge degradation. Further it comprises of three upstream and downstream-sections all sporting RELU activation. Each corresponding downstream section is linked by a residual connection to it's opposite. With a convolutional starting depth of 64 filters and bottleneck-layer of 32 by 32 by 512 it supports 15 convolutional hidden-layers as well as 3 max-pooling transformations for down sampling in it's practical configuration utilized in this work. Oppose to the downstream section the upstream employs regularization in form of a 20% dropout to prevent overfitting and other than sampling, convolutional up-scaling is applied. The input size is determined by the above mentioned histology data set [16] which is 256 by 256 by 3. The output layer uses Softmax with a categorical crossentropy loss function and an Adam-optimizer [20] and is of size 256 by 256 by 6. Weights are initialized by the method of He et al. [21]. As a batch size 18 was chosen and a training run took 60 epochs with an, every 10 epochs by a factor of $e^{-\frac{1}{10}}$ digressing, learning rate starting at 0.004. Experiments were conducted on Fold 2[3] of the data set, which provided a sufficient size of 2523 samples as well as fitting our hardware limitations. Model two, having the same parameters as model one, was trained with data perturbed by method one. Model three was trained by data augmented by method two also retaining parameters of model one. And finally, model four was trained by a combination of the three data sets with the same parameters as the other three models. Their outputs are evaluated against PGD with 100 steps, targeting the full area background class and a negative epsilon of 1‰. This is a very powerful attack as confirmed by Figures 3 and 8 as it makes the loss function step towards the provided full area target. Clearly expressed and fully visible nuclei are not recognized by the network. It should be noted that the base reading of the not-enhanced model one against the previously explained PGD attack is at an accuracy of around 80%. This is easily explained by the fact that the pixels of the five semantic classes together only hold 3.2% of the non-zero pixel values, while the background class's pixels are over proportionally filled with 81.6% of all pixels holding values above zero in this class. 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. The different loss values seem to be of significance, as they indicate a veritable transformation of the decision boundaries under the different training sets. All four models showed fluctuation in validation loss during the training of previous iterations which were leading to the results presented in this paper. They are documented in Figure 6. It is not entirely clear why this (besides maybe label noise and overfitting) happens, but the addition of hyper-parameter tuning routines in form of a digressing learning rate reduced the effects of this phenomena noticeably as shown in subsequent visualizations like Figure 7. The authors suspect that these dips stem from the network finding data representation in lower dimensional feature space thereby circumventing the curse of dimensionality [22]. This, once more indicates the need for an explainable feature curation. Another peace of circumstantial evidence is our observation of imprecision associated with the data labeling which could be causing said fluctuations in the fashion described before. A similar effect is also commented in the paper of Rolnick et al. [23]. They found that under heavy label noise lower learning rates tend to give better results.

TABLE 1
Test Results on 561 unperturbed and unseen Images

Model#	Loss	Accuracy	Dice Coef.	Jaccard Score
1	0.6718	0.8983	0.8946	0.8093
2	0.7276	0.8653	0.88	0.7853
3	0.6366	0.8955	0.8905	0.8026
4	0.6092	0.8917	0.883	0.7905

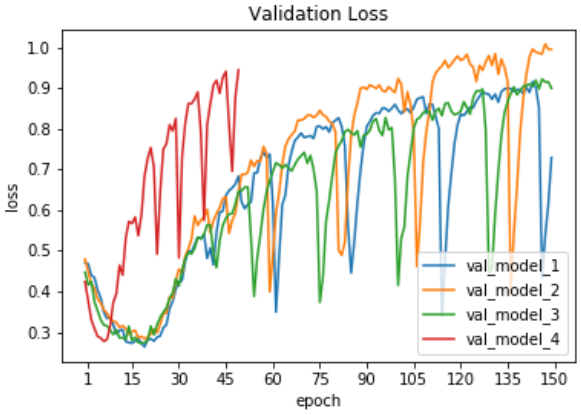

Fig. 6. Graph validation loss for all four Models of previous implementation without adaptive learning rate

In Table 2, we observe a basically unchanged accuracy for model four compared to model one in Table 1. An example of the generalization abilities of model four under the influence of perturbations is shown in Figure 11. Even though the complete training set for model 4 comprises of images perturbed by method one and two as well as the original training set, model four even surpasses model two, which was trained only on adversarial examples, by a slight margin under the regime of only perturbed data as depicted in Table 2. More so it surpasses the abilities of model two by far in an unperturbed setting as shown in table 1. This seems

2. Available at https://github.com/PonderWOnder/PokeNet
3. Available at https://warwick.ac.uk/fac/sci/dcs/research/tia/data/pannuke/fold_2.zip

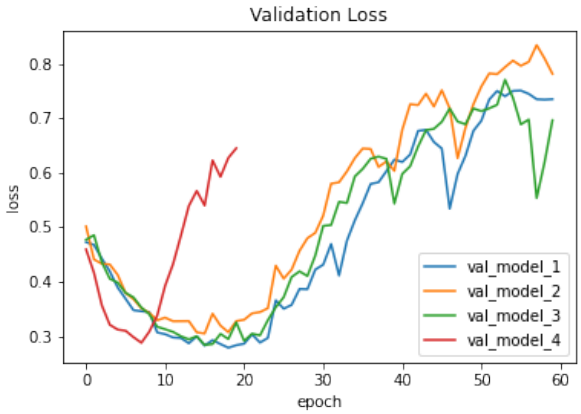

Fig. 7. Graph validation loss for all four Models in contemporary implementation with digressing learning rate

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 10 representing model three. Particularly in direct comparison with Figure 9 showing the output of model two, which was trained only on PGD perturbations.

TABLE 2
Test Results on 561 by 100 step PGD with Epsilon of 1‰perturbed and unseen Images

Model#	Loss	Accuracy	Dice Coef.	Jaccard Score
1	2.6005	0.8274	0.8303	0.71
2	0.7334	0.8852	0.88	0.7851
3	1.1432	0.8625	0.8593	0.7533
4	0.6398	0.8892	0.8806	0.7867

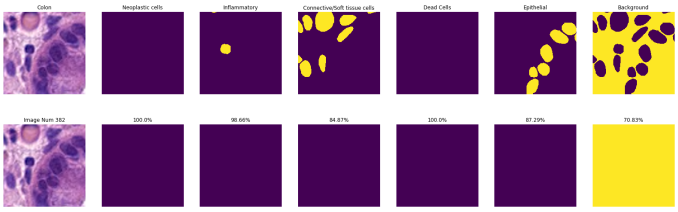

Fig. 8. Prediction of Model 1 against a 100 Step PGD with Epsilon of 1‰

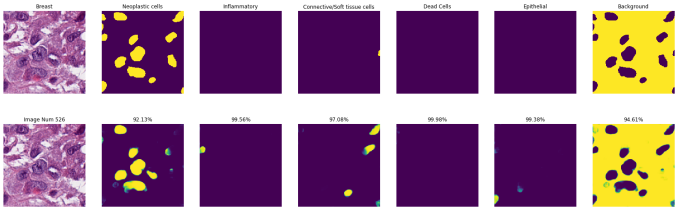

Fig. 9. Prediction of Model 2 against a 100 Step PGD with Epsilon of 1‰

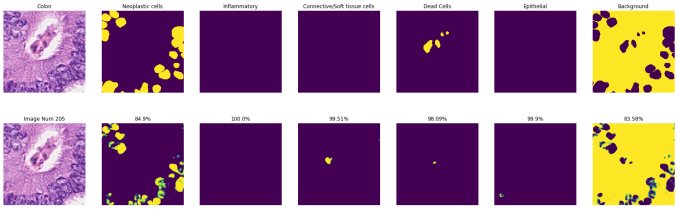

Fig. 10. Prediction of Model 3 against a 100 Step PGD with Epsilon of 1‰

4 Related Work

The next section shall discuss prior and inspiring work that led to this paper. Even thought two recent surveys [24] [25] contain most of the literature mentioned in the following in greater detail, they do not draw the conclusions influential to this work. The resemblance between the two makes it evident that even though they were written almost a year apart, their content doesn't seem to have changed much. As well, the main focus lies on classification tasks, illustrating further the lack of research done in the field of semantic segmentation. Especially in the aspect of their robustness and trustworthiness from an academic standpoint.

4.1 Rethinking Generalization

Zhang et al. [26] showed, in their exceptionally well received work, that neural networks are capable of fitting labels to every consistently altered, but semantically completely obliterated input they presented in their paper. Spanning from random labels in CIFAR10 and ImageNet [27] to Gaussian noise distributed with equal variance and mean as the mentioned data sets, neural networks provided convergence in the supervised training regimes, without any real world consistency regarding the input and their labels. They further investigated the effects of regularization techniques to better control overfitting with similar (not to say equal) results for training error, therefore proofing that neural networks are generally agnostic to all patterns they are presented with, no matter whether the user is aware of them or not. This implicates that there is the need for a selection process other than data curation by human sensory impressions.

In their short (however, for this work, influential) paper, Fischer et al. [28] provided a more tailored approach for semantic segmentation. They show that the prior knowledge of a class affiliation in an input makes it possible to only attack certain parts of an image to force segmenting classification to completely ignore otherwise unaltered resemblances to be invisible for a Fully Convolutional Networks (FCN)8, as described by Long et al. [29] trained on Cityscapes [7], leaving no other interpretation but that certain patterns within the data were more distinctive to the network than shape itself. This further implicates that there is the need for a more sophisticated selection process. Similarly, Xie et al. [30] introduced the Dense Adversary Generation (DAG) algorithm, iteratively transforming prediction in both detection as well as segmentation environments, such as the former mentioned FCN segmentation and Faster R-CNN [31] algorithm for detection and segmentation in a pixel to

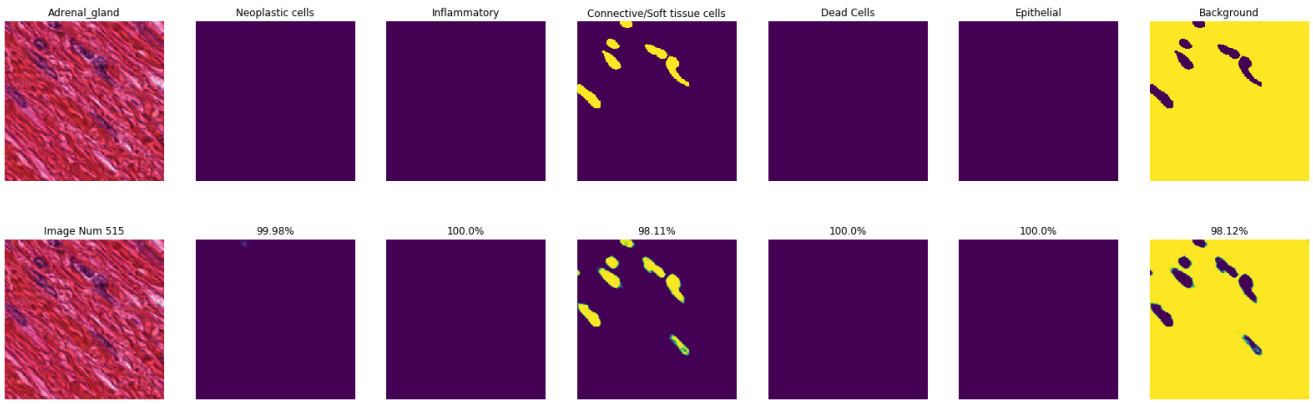

Fig. 11. Prediction of Model 4 against a 100 Step PGD with Epsilon of 1‰

pixel fashion towards a predefined target mask. Yet another attack was described by Sharif et al [32]. Their approach tricked Face Recognition Systems into misclassification, by adding strongly visible, but to certain regions of the image, restricted perturbations in form of glasses. This technique worked in both image and real world scenarios, and altered predictions by finding variance robust patterns that could be applied in variable conditions. Shortly after each other, Eykholt et al. [33] and Athalye et al. [34] showed that they can alter real world objects to provoke erroneous predictions by varying angle and distance, thereby bringing three-dimensional features as a further generalization problem in both classification and detection into play.

4.2 Adversarial Examples are not bugs

In their work, which has still not been reciprocated enough in our opinion, Ilyas et al. [13] proclaimed that "Adversarial vulnerability is a direct result of sensitivity to well-generalizing features in the data." To this intuition, they proposed a training set $\hat{D}_R$ from the representation layer of an adversarial trained network for the distribution D. Subsequent networks trained with distribution $\hat{D}_R$ showed adversarial robustness as well as comparable accuracy to a network trained by D in an adversarial regime. To set this into contrast, they also trained networks with random labels t. In this particular setup, they added small perturbations produced by the standard model to the images that pointed towards the random class t, thereby enforcing non-robust features that had adversarial character being the main indicator for the input's random label. This produced a new distribution $\hat{D}_{adv}$ comprising of x_{adv}, t, which, when used as training input for a new network, showed that features amplified by this technique seemed to suffice to give good accuracy for classifying samples of D, but was not able to withstand adversarial attacks showing even worst behavior than a network trained by the original distribution D.

Nakkiran [35] attempted to object by pointing out that adversarial examples are in fact bugs arising from mislabeled data that place decision boundaries within the reach of l_∞ and $l_{\{2}$ perturbations $< \epsilon$. He illustrated the effects that mislabeled data can have on a neural network, by drawing the decision boundaries of a binary classifier that had mislabeled data points while training. Ilyas et al. commenting Nakkiran's claims pointed out that their intent was not to negate the existence of noisy data, but to elaborate on the existence of two different types of features within the data that do not stand in contradiction with the phenomena description of Nakkiran.

In some sense, the findings of Ilyas et al. were confirmed in the work of Engstrom et al. [36]. They showed that at the representation layer neural network classifiers that were trained in a robust setting exhibit a far more human understandable feature extraction than non-robust networks do. More precisely, they showed that by applying a minimization problem for l_2 distance $x_1' = x_1 + \min_{\delta} \|R(x_1 + \delta) - R(x_2)\|_2$ to any input x_1 towards x_2 the representation resulting from $R(x_1')$ yields much more comprehensible visualisations, if $R(x)$ was trained towards robust classification. This further illustrates that feature selection is indeed a vital tool for semantic coherent representations.

4.3 Adversarial Defense

In their study on the robustness of semantic segmentation, Arnab et al. [18], demonstrated different attack vectors, similar to the ones used in our work, based on FGSM, using single step as well as multi-step whitebox attacks against commonly used semantic segmentation setups both residual and sequential. Their findings show that residual based networks, like for example E-Nets [37], show greater robustness after training with perturbed inputs than the sequential VGG [38] based ones. They conclude that in safety critical conditions it is advisable to chose frameworks that utilize multiscale processing, since these present themselves with greater robustness over accuracy after adversarial training [12]. In [17], Tsipras et al. present a very compelling argument that every classifier that reaches an accuracy of $1 - \delta$ will have an adversarial accuracy of $\frac{p}{1-p}\delta$, where p is the $p < 1$ correlation of the label with an input feature x_1 and $\delta \to 0$. Thus, reaching a high standard accuracy relies on brittle features that exist in $[0, 1 - p_{x_1}] := \{p_{x_d} \in \mathbb{R} | 0 < p_{x_d} < 1 - p_{x_1}\}$ – which in turn are easy to attack.

5 Discussion and Conclusion

Although network four presents robustness and a little better accuracy than the adversarially trained model two when

presented with perturbed inputs, 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 but common metrics like Jaccard Score and Dice Coefficient are listed too in tables 1 and 2. There may be better ways to account for mutliclass-errors, but due computational-limitations we used the mean point for point accuracy for optimization. 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 [39] for subsequent research since they provided an indisputable ground-truth, even though we provide a rudiment patter generator with in the code which showed similar results as real-world data in terms of robustness and accuracy increase but was not able to reproduce complexity present in naturally aqquired distributions. Nonetheless, we showed that method two improved behaviour of adversarial training [11] when data augmented by this process is added into the training distribution and therefor is worth a more rigorous investigation.

We were not able to integrate code that could document the high dimensional transformation applied to the inputs in a way that would have any understandable meaning to humans due to a lack of explainablity concepts for semantic segmentation, similar to [36] framework for classification, that would facilitate such a task. In our opinion, this is an interesting topic for further investigation since such transformations 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 and strongly fluctuating loss values for the for the same input to all four models.

Furthermore we emphasise that the inherent brokenness of all CNNs, stemming from a well defined continuous derivable function at the heart of the mechanism, will not be solved by adding perturbations [40] into training routines, transforming inputs [41] until perturbations lose their effectiveness nor obfuscating gradients [42] produced by derivatives of such functions 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.

Acknowledgments

This research was funded by the European Union's Horizon 2020 research and innovation program under grant agreement No.826078 (Project 'FeatureCloud'), 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] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[2] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and 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.

[3] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2020, pp. 1–16.

[4] C. Kamann and C. Rother, "Benchmarking the robustness of semantic segmentation models," *ArXiv*, vol. abs/1908.05005, 2019.

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

[6] X. Liu and C.-J. Hsieh, "Rob-gan: Generator, discriminator, and adversarial attacker," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016. [Online]. Available: http://arxiv.org/abs/1604.01685

[8] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *Courant Institute New York University*, 2010.

[9] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *University of Toronto*, 2009.

[10] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *stat*, vol. 1050, p. 20, 2015.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

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

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

[15] O. Ronneberger, P. Fischer, and 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.

[16] J. Gamper, N. A. Koohbanani, K. Benet, A. Khuram, and N. Rajpoot, "Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification," in *European Congress on Digital Pathology*. Springer, 2019, pp. 11–19.

[17] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," *arXiv preprint arXiv:1805.12152*, 2018.

[18] A. Arnab, O. Miksik, and P. H. Torr, "On the robustness of semantic segmentation models to adversarial attacks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[19] P. Nakkiran, "Adversarial robustness may be at odds with simplicity," *arXiv preprint arXiv:1901.00532*, 2019.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[22] R. Bellman, R. Corporation, and K. M. R. Collection, *Dynamic Programming*, ser. Rand Corporation research study. Princeton University Press, 1957. [Online]. Available: https://books.google.at/books?id=wdtoPwAACAAJ
[23] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," *arXiv preprint arXiv:1705.10694*, 2017.
[24] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.
[25] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346 – 360, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S209580991930503X
[26] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.
[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
[28] V. Fischer, M. C. Kumar, J. H. Metzen, and T. Brox, "Adversarial examples for semantic image segmentation," *arXiv preprint arXiv:1703.01101*, 2017.
[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
[30] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1369–1378.
[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015, pp. 91–99. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf
[32] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1528–1540. [Online]. Available: https://doi.org/10.1145/2976749.2978392
[33] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.
[34] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 284–293.
[35] P. Nakkiran, "A discussion of'adversarial examples are not bugs, they are features': Adversarial examples are just bugs, too," *Distill*, vol. 4, no. 8, pp. e00 019–5, 2019.
[36] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, and A. Madry, "Adversarial robustness as a prior for learned representations," *arXiv preprint arXiv:1906.00945*, 2019.
[37] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," 2016.
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[39] A. Holzinger, M. Kickmeier-Rust, and 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.
[40] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 3358–3369. [Online]. Available: http://papers.nips.cc/paper/8597-adversarial-training-for-free.pdf
[41] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, "Countering adversarial images using input transformations," *arXiv preprint arXiv:1711.00117*, 2017.
[42] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," *arXiv preprint arXiv:1802.00420*, 2018.

Konstantin Avramidis performed this work during his summer internship at the Medical University Graz in the Human-Centered AI Lab of Professor Holzinger. Konstantin is currently a student at the IT Department of the University of Applied Sciences, St.Pölten, Austria. His research interests are in machine learning and particularly in robustness.

Peter Kieseberg (M'11–SM'18) heads the Institute of IT Security Research at the St. Pölten University of Applied Sciences and the Josef Ressel Center for Blockchain Technologies & Security Management. Peter's research interests mainly focus on issues surrounding privacy and data protection in machine learning and data driven environments.

Rudolf Mayer is a senior researcher and lead of the machine learning and data management team at SBA Research, Vienna, Austria, and a lecturer at Vienna University of Technology. His research interests include information retrieval (focusing on text and music data), and machine learning. Specifically, he focuses on privacy-preserving data publishing and machine learning, as well as security aspects of machine learning (adversarial machine learning).

Edgar Weippl is professor at the University of Vienna and scientific director of SBA Research. In 2004 he joined the TU Wien and founded the research center SBA Research together with A Min Tjoa and Markus Klemen. In 2020 Edgar left TU Wien to accept a position as full professor at the University of Vienna, Faculty of Computer Science. Edgar's research focuses on fundamental and applied research in security, as well as blockchains and security of production systems engineering.

Andreas Holzinger (M'00) is Visiting Professor for explainable AI at the Alberta Machine Intelligence Institute of the University of Alberta, Canada since 2019 and head of the Human-Centered AI Lab at the Medical University Graz, Austria. He received his PhD in cognitive science from Graz University in 1998 and his second PhD in computer science from Graz University of Technology in 2003. He is IEEE Member since 2000.