The Role of Input Gradients as Indicators for Discriminative Features

Felix Bötte
felix.boette@student.uni-tuebingen.de

Felix Pfeiffer
f.pfeiffer@student.uni-tuebingen.de

Abstract—**Gradient-based methods are a popular method to give interpretable explanations to the underlying decision processes of a Neural Network. Generally, input gradients are a good indicator on which features a Neural Network considers more important, and which features have less influence on the prediction. However, the classification boundaries of gradient-based explanation methods can be unprecise which leads to falsely highlighting gradients of certain features. In this paper, we investigate the role input gradients play in identifying discriminatory features and whether those are able to distinguish between different classes. First, the paper discusses literature on this topic. Different methods, like visualization of the gradients by saliency maps, showed that input gradients of robust networks are better interpretable because they mainly highlight only the relevant input features, compared to standard networks. Competing results were obtained by regularizing the gradients. In the second part of the paper we investigated the effects of regularizing gradients and weights on the interpretability of the input gradients experimentally with a toy dataset. We came to the conclusion that regularizing the weights with ℓ_2-loss is the most successful way in terms of obtaining interpretable input gradients that mostly highlight discriminative features. We deduce that in general the role of input gradients for discriminative features is not clearly definable. However, for robust networks it is likely to infer discriminative features from the input gradients.**

I. Introduction

Nowadays, Machine Learning (ML) has gained great importance in many aspects of our life. This ranges from improving the efficiency of industrial plants to personalized recommendations in online stores or streaming services to improve medical care by supporting diagnosis and treatment decisions [1]. ML has the potential to relieve people of work or simplify work in many aspects of life. However, it is important to develop explainable and fair machine learning (ML) methods for various reasons. E.g. to ensure that the decisions the models make are reasonable and fair, build people's confidence in these models, identify and correct errors in ML models, reduce biases and uncertainties in ML systems and many more. This is especially important when ML models are used in sensitive areas such as credit scoring, insurance quotes, or healthcare. The past has shown that people have been wrongly denied parole because of an ML model which has disadvantaged people because of bias in the data used to build this model [2].

One of the main problems in this domain is the uncertainty which aspects or features of an input have which influence on the decision of a Neural Network (NN). Especially decisions of Deep Neural Networks (DNNs), which consist of several hundred or thousand hidden units and have several hundred million parameters, are very difficult to interpret or even uninterpretable. There exist many different approaches to analyze which features are most important for an ML model and how they contribute to the decision making process. One way to to examine the relevancy of certain input features for the prediction of an ML model are so called gradient-based methods. These methods are based on the concept of gradient descent, which is one of the most widely used technique to minimize the error rate of a NN by adjusting the weights and biases [3]. By measuring the intensities of those gradients one can get an idea how strong the algorithm values a certain input in the final decision process. However, these techniques are often not robust to perturbations of the input data [4].

A. Motivation

In this paper, we aim to explore the relationship between input gradients and discriminative features in NNs. Thus, we have searched the literature for explanations and hypotheses on the role of input gradients as indicators for discriminant features and which factors are found to be influential. Furthermore, we were looking for evidence that suggests that adversarial robust training of DNNs is linked to more robust gradient interpretation and consequently better explainability of the underlying prediction mechanism.

In order to investigate this topic empirically we use a toy dataset provided by [5] to show whether there is a relationship between the regularization of gradients and the regularization of weights with ℓ_1, ℓ_2 and ℓ_∞ loss, or whether the application of these two methods equally lead to more interpretable NNs.

B. Paper Structure

In section two we summarize findings in the literature.

In section three we present our experiments we have done based on the dataset and previous work from [5] and discuss our findings.

Finally, we give a brief summary in section four.

II. Findings in the Literature

The authors of [5] make the assumption (A): *"Coordinates with larger input gradient magnitude are more relevant for model prediction compared to coordinates with smaller input gradient magnitude."*, and verify it with a custom evaluation framework they call `DiffROAR` and a data set they call

BlockMNIST. Additionally, they prove theoretically that (A) is not valid for standard one-hidden-layer multilayer perceptrons (MLPs). For this particular case with this particular data set and model, input gradients do not highlight specific input features. Subsequently, the authors trained ℓ_∞ and ℓ_2 robust models using Projected Gradient Descent (PGD) [6] adversarial training such that the adversarial examples stay within ℓ_∞ and ℓ_2-distance of ϵ from the original image.

Both, the standard and robust models are trained on FashionMNIST, ImageNet-10 and CIFAR-10 datasets. With DiffROAR, which builds upon the remove-and-retrain (ROAR) methodology [7] they tested whether feature attribution methods through gradient intensity satisfy (A) on realistic datasets. Input coordinates with higher attribution rank are supposed to be more important for model prediction than ones with lower rank. The results show that models that are robust to ℓ_∞ and ℓ_2 perturbations satisfy (A) better for every percentage of unmasked pixels $k < 100\%$. Standard MLPs show no better than model-agnostic random attributions. Furthermore, for $k < 40\%$, standard ResNet models trained on CIFAR-10 and ImageNet-10 strongly violate (A), giving evidence that coordinates with top-most gradient attribution rank have worse predictive value than coordinates with bottom-most rank. In contrast, robustly trained ResNet models satisfy (A) consistently.

In a second experiment the authors investigate (A) on the BlockMNIST dataset for standard and robust models. The dataset consists of a combined image block where one part is a MNIST image functioning as a *signal* block and a uniform *null* block. The non-discriminative null block contains, in contrary to the signal block, no information about the class. The signal and null blocks were stacked in a random order on top of each other. The results showed that standard MLP and ResNet models highlight the signal block as well as the non-discriminative null block whereas robust models only highlight the signal block and therefore satisfy (A).

As a possible explanation for those findings the authors hypothesize that when discriminative features vary across instances (e.g., signal block at top vs. bottom), input gradients of standard models not only highlight instance-specific features but also leak discriminative features from other instances. The authors were referring to this as feature leakage.

Using DiffROAR framework, the before mentioned empirical analysis on BlockMNIST and a theoretical examination, the authors present strong indication that that standard models do not satisfy (A). In contrast, adversarially robust models satisfy (A) consistently.

The researchers of [8] have been investigating the method of regularizing the gradient norm of the output of a neural network model in contrast to use adversarial input examples to increase classification robustness and accuracy. The idea of regularizing a model's output gradient norm goes back to the idea of *Double Backpropagation* [9]. Generally, gradient regularization penalizes large output gradients in order to accomplish smooth priors. The authors argue that this regularization approach increases a models classification accuracy, especially on small datasets. In their experiments the authors compared gradient-based regularization with mutiple other methods i.e. *Jacobian Regularization* [10]. They trained a ResNet model on a reduced version of CIFAR10, ImageNet and MNIST datasets whith no data augmentation being applied. The results indicate that gradient regularization increases classification accuracy and even more when additionally combined with weight regularization. As a noteworthy objection to this method the authors mention that it the gradients are only regularized in the training data. Nonetheless, gradient regularization models generally become smoother on the whole data manifold where gradients tend to be more globally controlled and thus be more robust and interpretable.

The authors of [11] hypothesize that by specifically training DNN models to have smooth, non-vanishing input gradients with fewer extreme outliers, it will be more interpretable as well as more resistant to adversarial examples.

Ross et al. are confirming their hypothesis by using gradient regularization during the training process on MNIST, notMNIST and SVHN of their models and comparing them to adversarially trained and defensively destilled models using softmax [12]. The results show that gradient regularization leads to robustness towards other models' adversarial examples at high perturbation rate, while all other models are fooled by gradient-regularized model examples. Furthermore, a human subject study was conducted in order to evaluate the quality of the adversarial examples and gradient interpretability. The findings suggest that the adversarial examples by gradient-regularized models are most convincing and best classified as their target. When mispredictions happen they were also rated as the model with the most 'reasonable' misclassification. Moreover, the gradient visualization of gradient regularized models is more intuitively aligned with human reasoning when it comes to highlighting discriminative features.

Summarized, the results indicate that when comparing the regularization method to adversarial robust training it yields equally or better results in terms of robustness and interpretability. Those two methods can be combined to achieve even greater robustness. Furthermore, gradient regularization significantly shapes the shape of the decision boundary which suggests that the model makes predictions for qualitatively different reasons. However, the increased robustness comes at the price of increased training time since gradient regularization is a second order method which increases training time slightly more than by a factor of 2.

The team behind [13] investigated the differences between gradient regularization and adversarial training as well and show experimentally that input gradient regularization is competitively comparable without leading to gradient obfuscation. This hypothesis is based on the idea that small loss gradients should lead to more robustness against gradient based attacks since perturbation in the input does not lead to high magnitude changes in the resulting gradient. They do so by training their model with ℓ_∞ and ℓ_2 gradient regularization. Furthermore, they show theoretically that bounding the minimum adversarial distance to lower bounds on the minimum perturbation intensity necessary to make wrong classifications can be competitively compared to state-of-the-art attacks. The results suggest that regularization with squared ℓ_2 norm is more robust than adversarial robust models on `CIFAR10` and `ImageNet` datasets.

The authors of [14] also try to answer the question of how adversarial robustness and gradient interpretability are related. For doing so, they hypothesize that the reason why loss gradients from adversarial robust models align better with human perception might be that robust training restricts gradients closer to the image manifold. When comparing standard models to adversarial robust models the results indicate that with increasing strength of attack, increasing ϵ-value, the closer the prediction lies to the image manifold of the test data. Consequently, it aligns more with human reasoning and being more interpretable. As a possible explanation for this phenomenon the authors refer to the boundary tilting perspective [15] which states that adversarially trained models remove tilting along directions of low variance in the data and thus creating robust decision boundaries. Only strong attacks, which large ϵ-value would be able to significantly change the classification process. The authors confirmed their hypothesis by training three two-layer ReLU networks on a 2-dimensional toy dataset to classify points from two distinct bivariate gaussian distributions. The first network is trained on original data and the second and third network are trained on weak and strong adverseries respectively. The results show that the decision boundary of the standard network is tilted along the direction of low variance in data and training against increasingly stronger adverseries removes the tilt to larger degree , thus adversarial examples align better with data manifold.

In order to show that that adversarial robustness is linked to better interpretability the authors conducted another study using the Remove and Retain `ROAR` and Keep and Retain `KAR` framework [16]. ROAR replacing the most important and KAR the least important pixels of an input image, a better attribution method should cause more or less accurate degradation respectively. Results show that there is a strong correlation between the strength of an attack and interpretability using the ROAR and KAR methodology.

However, the authors also show empirically that there is a trade-off between test accuracy and loss gradient interpretability as suggested by previous studies [17] [18]. In order to verify this hypothesis the author again trained CNNs on `CIFAR10` under various adversarial attack settings and showed that is a near-monotonic decreasing relation between interpretability and accuracy under both ROAR and KAR framework. They also observed that ℓ_2-trained networks are more robust to this trade-off than ℓ_∞-trained networks in ROAR and the other way around for KAR. This suggests that attributions from ℓ_2-trained networks are better at highlighting important features and ℓ_∞-trained networks are better at identifying less-important features. Lastly, the authors provide possible solutions to this trade-off, namely combining adversarial training with global attribution methods like Integrated Gradient [19] and seeking better ways to utilize ℓ_∞-training since there is a large performance gap to ℓ_2-training but a less strong robustness difference.

The team behind [20] investigated the correlation between smooth decision boundaries and increased interpretability as well. They hypothesize that when this is the case the smooth input gradients of a DNN model will more closely align with the normal vectors of adjacent boundaries. Consequently, as already mentioned in the previous part, adversarial robust models have smoother boundaries and therefore enable gradient-based attribution methods like Integrated Gradients to identify more accurate estimations about nearby decision boundaries. This also results in better visual interpretability since there is a more clear separation of gradients and decision boundaries. The results by the authors give a geometrical explanations for the correlation between smooth gradients and interpretability and align with findings from Ilyas et al. [21] who hypothesise that robust models learn robust and relevant features.

Nonetheless, the authors mention that their idea of using bounding boxes for interpretability has limitations since they are not perfect ground-truth knowledge for attributions. There exist multiple cases where the boundaries are to big or fail to incorporate all relevant features.

The authors of [21] also examined the relationship between the distance to a decision boundary and the alignment with the right class allocation. They hypothesize that with increasing distance to a nearby decision boundary the greater is the alignment which leads to more interpretable saliency maps for what the model has learned. However, it is important to mention that the results from the authors only apply for linear models. For validating their hypothesis they trained multiple models with 1000 different adversarial examples on both `MNIST` and `ImageNet` using double backpropagation to increase robustness [22]. The results indicate that the connection between alignment and interpretability is greatly connected to how similar the neural network is to a linear model locally. This became apparent for the evaluation on `ImageNet` where accurate models tend to be more non-linear than on comparatively simpler problems like `MNIST` dataset.

The findings of [23] align with the results from [24] mentioned before, as they agree that the distance of samples to the decision boundary is linked to model robustness and interpretability. Furthermore, they show that DNNs have high invariance to non-discrimnative features and highlight that the decision boundaries can only exist when the corresponding classifier is trained with certain adversarial examples that hold them together. These boundaries are very sensitive to the exact shape of the training data which confirms the hypothesis that adversarial training helps building more robust classifiers by strengthening the shape of the classification boundary. More precisely, the data points that strengthen the boundaries the most are the ones closest to them.

In [25], the writers address the question why gradients in standard models are highly structured and explanatory when the can actually be arbitrarily manipulated. They also address the question of what aspect the models input-gradients depend if they do not depend strongly on the underlying discriminative function. The softmax function is defined as follows:

$$\sigma(z) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } j = 1, \ldots, K \tag{1}$$

The vector z contains the so-called logits which are predictions made by the model to rate the assignment of a certain input to one of the K classes. The softmax function calculates a probability distribution where all class predictions of the model add up to one. It makes no difference if some arbitrary but constant function g is added to every logit j. Therefore, the logit-gradients can be arbitrary without changing the underlying discriminative function of the model. The logits are interpreted as unnormalized log-densities of a class-conditional implicit density model by the authors. Srinivas and Fleuret hypotize that: ”Perhaps input-gradients are highly structured because the implicit density model is aligned with the ground truth class-conditional data distribution.” In order to validate this hypothesis they introduce a method called score-matching to increase and decrease the alignment between the implicit density model and the data distribution. Experiments were performed on the `CIFAR10` as well as `CIFAR100` dataset using a ResNet18 model. The NNs were trained with score-matching and anti-score-matching: first when the alignment between the implicit density model and the ground truth increased, and secondly when it decreased. In addition, the the authors trained a NN with gradient-norm regularization. By using pixel perturbation, it was found that the NN trained with score-matching and gradient-norm regularization detect less relevant pixels better than the standard model as well as models trained with anti-score-matching and are consequently more robust. Moreover, the saliency maps of the score-matched and gradient-norm regularized NNs were more perceptually aligned with the input data, allowing for better interpretation of the input gradients. Finally, the authors fail to answer why the implicit density models of pre-trained NNs are better aligned with the ground truth data. However, they indicate that the presence of an implicit gradient-norm regularizer in standard SGD could be a possible cause.

III. Experiments

In order to further investigate the influence of ℓ_∞ and ℓ_2 gradient as well as ℓ_1 and ℓ_2 weight regularization on the interpretability of NNs we add additional experiments to the toy dataset from [5]. Our goal is to show in which way different regularization techniques during training of neural networks effect the level of highlighting discriminative features without being easily perturbed by varying input features in their classification process. Therefore, we are trying to show that gradient and weight regularization on one-hidden-layer MLPs trained on the toy dataset only highlight instance-specific coordinates as stated in assumption (A) by the authors [5].

A. Dataset

The dataset we use is comprised of a number of data points where the label $y = \pm 1$ with probability 0.5 and $x = y \cdot e_j$ where j is chosen uniformly at random from $1, \ldots, d$ where d denotes to the number of blocks as described in [5].

The idea is to create a blocks that represent task-relevant signal blocks that is informative of the target label. Additionally, there are noise blocks that do not contain task-relevant signals for any input instance. Consequently, this experiment setup should help visualizing if regularization techniques during training influence discriminative power of the corresponding models. The resulting models should only highlight task-relevant signal blocks through significant gradient exposure.

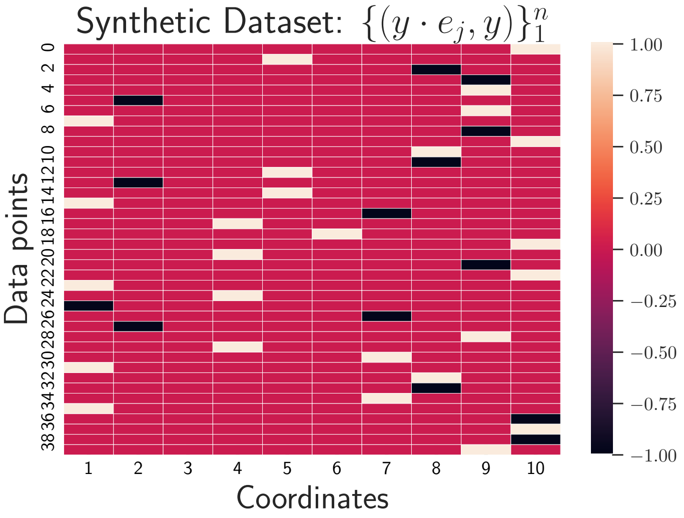

Fig. 1. The first 40 datapoints of the testset we used for our experiments.

B. Methods

For our regularization methods we use gradient as well as weight regularization. More precisely, we use the ℓ_1, ℓ_2 and ℓ_∞ norm to restrict the input gradient and also use ℓ_1, ℓ_2 and ℓ_∞ norm to regularize the arbitrary growth of weight values. Furthermore, we investigate whether the combination of gradient and weight regularization leads to even more robustness and discrimnative capability. Consequently, we trained 11 different models: A standard model without gradient and weight regularization and the same model with ℓ_1, ℓ_2 and ell_∞ weight regularization respectively. And a ℓ_1, ℓ_2 and ℓ_∞ gradient regularized model with and without the two types of weight regularization. Furthermore, we combined ℓ_1 and ℓ_2 weight and gradient regularisation such as regularize both weights and gradients with the same loss. Additionally, we also considered adversarial attack methods of the listed regularizers above to investigate the relationship between gradient regularization and corresponding attacks.

C. Models

As a model to test our hypothesis we consider a one-hidden layer neural network with ReLU as activation function. The hidden layer has a width of 25000 nodes and as a learning rate μ a factor of 0.1 has been chosen. Additionally, a decay value of 0.75 has been added every 50 epochs to a minimum of 0.001

All model variants have been trained for 4000 epochs and achieved 100% test accuracy on the linear separable synthetic dataset.

D. Results

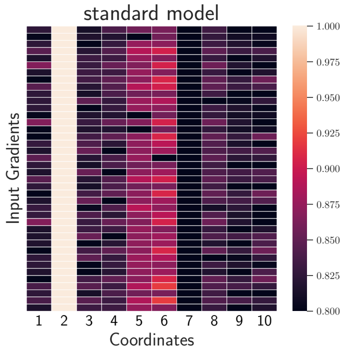

Fig. 2. Size of the normalized gradients on the test data of a standard model without regularization.

In figure 2 you can see that the gradients of the standard model are not very well interpretable. For example, the gradients on feature 2 are always significantly larger than the gradients on the other features. Only the data points 2, 12 and 14 really contain information about the label of the input data.

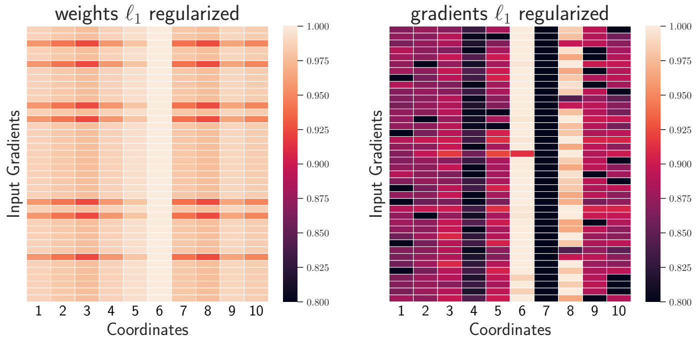

Fig. 3. Size of the normalized gradients on the test data of a model trained with ℓ_1 weight and ℓ_1 gradient regularization.

Figure 3 shows the gradients of a model in which the weights were regularized with the ℓ_1-loss during training, as well as a model in which the gradients were regularized with the ℓ_1-loss. It is very noticeable that the gradients of the weight regularized model after normalization are all about the same size. This means that the pattern of the test data set from figure 1 can not be found in the gradients, which makes it difficult to interpret the decision-making of the NN based on the gradients. The regularization of the weights also contributes only marginally to a better interpretability. For example, for feature 8, the pattern of the test data can be found in the gradients. However, this does not apply to the other features.

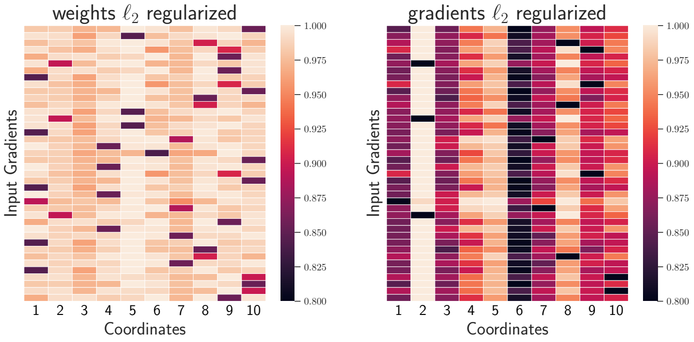

Fig. 4. Size of the normalized gradients on the test data of a model trained with ℓ_2 weight and ℓ_2 gradient regularization.

The best visual result was achieved in figure 4 by regularizing the weights with the ℓ_2-loss. One can clearly see that only at the coordinates where the information is located, the gradients stand out clearly from all other gradients. In a similar but weakened form, this also applies to the model with ℓ_2 regularized gradients. However, the result is not as clear because the gradients for e.g. feature 1 and 6 are always generally smaller, even if there is no information there.

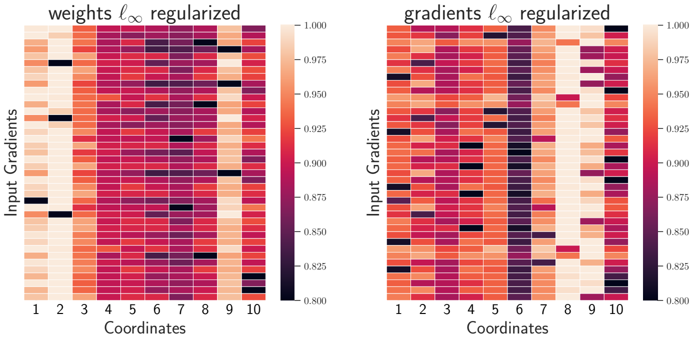

Fig. 5. Size of the normalized gradients on the test data of a model trained with ℓ_∞ weight and ℓ_∞ gradient regularization.

By regularizing the weights with ℓ_∞-loss, the gradients at the coordinates, where information is contained, stand out. However, with feature 1 you can see in figure 5 that not all coordinates are noticeable. For example, no particular gradient is evident for data points 31 and 35. Regularizing the gradients with ℓ_∞-loss results in all important coordinates being clearly visible. Nevertheless, general differences in the gradients among the various features are also detectable.

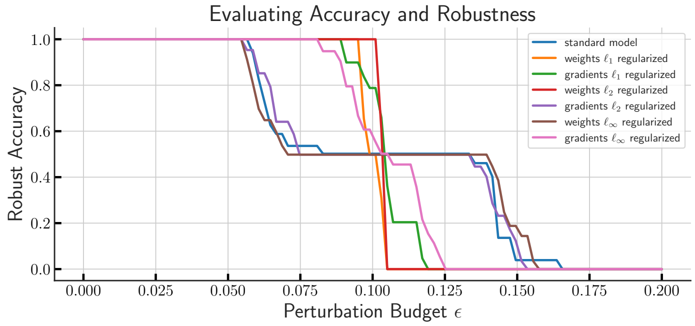

Fig. 6. The accuracy on the test set against the strength ϵ of the adversarial attacks is plotted. This allows to visualize the robustness of the individual models.

The plot shows that the standard, the ℓ_2 gradient regularized and the ℓ_∞ weight regularized models are the least robust to changes in the input data. This matches relatively well with the visualizations of the gradients. However, it is astonishing that the regularization of the weights as well as the gradients with the ℓ_1-loss lead to more robust models although the gradients visually fit much worse to the input data. The model where the weights are regulated with ℓ_2 during training is the most robust when it comes to 100% accuracy. This is consistent with the findings from figure 4. The fact that the standard model l2 gradient regularized and ℓ_∞ weights regularised model are at exactly 50% accuracy above a certain perturbation budget is probably due to the fact that it is a binary classification problem.

IV. Discussion

In this paper, we conducted a meta study on the role of input gradients for highlighting discriminative features. In order to tackle this problem we mainly found three different approaches: regulating the gradients, the weights or use adversarial training. Finally, the collective conclusion of most of the paper referenced is that the gradients of adversarial robust NNs are more interpretable than those of standard models. The same applies for models where the gradient is regularized during the training process. Another common finding is that the decision boundary of robust models changes in such a way that the model can make qualitatively better decisions which features are important. This is because the different classification boundaries are further apart on the image manifold, meaning the variance between the individual classes is increased. The resulting structure leads to mostly highlighting only discriminative features and less cases where the model misclassifies the importance of a feature due to overlapping decision boundaries. Having a more clearly separated decision boundary for when a certain feature is important to a certain class consequently leads to better interpretable results that align with a human way of reasoning.

Our experiments suggest that for the dataset and architecture used, regularization of the weights leads to more robust results than regularization of the gradients. Moreover, we achieved the best results in terms of robustness and interpretability with the ℓ_2 weight penalty. A possible explanation for this could be that with ℓ_1 regularization the penalty is proportional to the absolute value. Therefore, the regularization leads to the case that only a subset of all available features is used for the decision process since many weights or gradients tend to take the value 0. The figure showing the gradients of the ℓ_1-regularized weights highlights that all gradients have the tendency to converge to more or less the same value. Since ell_2 is proportional to the square of the weights or gradients, they are only scaled down, but never reach the value 0. This possibly leads to a better generalization and therefore to a more robust and interpretable model. On the other hand, ℓ_∞ regularization penalizes only the largest value, constraining the weights and gradients to a specific range. This may not be greatly useful for explainability because either not all weights and gradients are penalized, or only important weights and gradients are penalized, even if they are not supposed to.

By regularizing gradients and weights we did not gain any new insights. For example, regularizing both, the weights and gradients, with ℓ_2 norm made no big effect on the result when compared to only regularized weights. When we tried other combinations of weight and gradient regularization the results were always similar to the results when only weights or gradients were regularized individually. To put it in a nutshell, appropriately scaling down the decision boundary of a NN model through gradient and weight regularization greatly influences its robustness to adversarial attacks as well as helping to omit less informative features in the underlying classification process. Both helps with the models interpretability to align more with human reasoning and robustness towards changes of unimportant features in the input space.

References

[1] Kononenko, Igor. ”Machine learning for medical diagnosis: history, state of the art and perspective.” Artificial Intelligence in medicine 23.1 (2001): 89-109.

[2] Hao, Karen. “AI is sending people to jail—and getting it wrong.“ Technology Review 21 (2019).

[3] Ghorbani, Amirata, Abubakar Abid, and James Zou. ”Interpretation of neural networks is fragile.” Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

[4] Ruder, Sebastian. “An overview of gradient descent optimization algorithms.“ arXiv preprint arXiv:1609.04747 (2016).

[5] Shah, Harshay, Prateek Jain, and Praneeth Netrapalli. “Do Input Gradients Highlight Discriminative Features?.“ Advances in Neural Information Processing Systems 34 (2021): 2046-2059.

[6] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In ICLR Workshop, 2016.

[7] Sara Hooker, D. Erhan, P. Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In NeurIPS, 2019.

[8] Varga, Dániel, Adrián Csiszárik, and Zsolt Zombori. ”Gradient regularization improves accuracy of discriminative models.” arXiv preprint arXiv:1712.09936 (2017).

[9] H. Drucker and Y LeCun. Double backpropagation: Increasing generalization performance. In Proceedings of the International Joint Conference on Neural Networks, volume 2, pp. 145–150, Seattle, WA, July 1991. IEEE Press.

[10] Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. IEEE Trans. Signal Processing, 65(16):4265–4280, 2017. doi: 10.1109/TSP. 2017.2708039. URL https://doi.org/10.1109/TSP.2017.2708039.

[11] Ross, Andrew, and Finale Doshi-Velez. “Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients.“ Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

[12] Ba, J., and Caruana, R. 2014. Do deep nets really need to be deep? In Advances in neural information processing systems, 2654–2662

[13] Finlay, Chris, and Adam M. Oberman. ”Scaleable input gradient regularization for adversarial robustness.” arXiv preprint arXiv:1905.11468 (2019).

[14] Kim, Beomsu, Junghoon Seo, and Taegyun Jeon. “Bridging adversarial robustness and gradient interpretability.“ arXiv preprint arXiv:1903.11626 (2019).

[15] Thomas Tanay and Lewis Griffin. A boundary tilting perspective on the phenomenon of adversarial examples. arXiv preprint arXiv:1608.07690, 2016.

[16] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. Evaluating feature importance estimates. In ICML Workshop on Human Interpretability in Machine Learning, 2018.

[17] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In International Conference on Learning Representations, 2018.

[18] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy? – a comprehensive study of robustness of 18 deep image classification models. In ECCV, 2018.

[19] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In International Conference on Machine Learning, 2017.

[20] Wang, Zifan, Matt Fredrikson, and Anupam Datta. ”Robust models are more interpretable because attributions look normal.” arXiv preprint arXiv:2103.11257 (2021).

[21] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In Advances in Neural Information Processing Systems, 2019.

[22] Etmann, Christian, et al. ”On the connection between adversarial robustness and saliency map interpretability.” arXiv preprint arXiv:1905.04172 (2019).

[23] Ortiz-Jimenez, Guillermo, et al. ”Hold me tight! Influence of discriminative features on deep network boundaries.” Advances in Neural Information Processing Systems 33 (2020): 2935-2946.

[24] Simon-Gabriel, C.-J., Ollivier, Y., Scholkopf, B., Bottou, L.,a nd Lopez-Paz, D. Adversarial vulnerability of neural networks increases with input dimension. arXiv preprint arXiv:1802.01421, 2018

[25] Srinivas, Suraj, and François Fleuret. “Rethinking the role of gradient-based attribution methods for model interpretability.“ arXiv preprint arXiv:2006.09128 (2020).