On the Connection Between Adversarial Robustness and Saliency Map Interpretability

Christian Etmann [*1,2] **Sebastian Lunz** [*3] **Peter Maass** [1] **Carola-Bibiane Schönlieb** [3]

Abstract

Recent studies on the adversarial vulnerability of neural networks have shown that models trained to be more robust to adversarial attacks exhibit more interpretable saliency maps than their non-robust counterparts. We aim to quantify this behavior by considering the alignment between input image and saliency map. We hypothesize that as the distance to the decision boundary grows, so does the alignment. This connection is strictly true in the case of linear models. We confirm these theoretical findings with experiments based on models trained with a local Lipschitz regularization and identify where the non-linear nature of neural networks weakens the relation.

1. Introduction

Despite impressive results in a variety of classification tasks (LeCun et al., 2015), even highly accurate neural network classifiers are plagued by a vulnerability to so-called *adversarial perturbations* (Szegedy et al., 2014). These adversarial perturbations are small, often visually imperceptible perturbations to the network's input, which however result in the network's classification decision being changed. Such vulnerabilities may pose a threat to real-world deployments of automated recognition systems, especially in security-critical applications such as autonomous driving or banking. This has sparked a large number of publications related to both the creation of adversarial attacks (Goodfellow et al., 2014; Kurakin et al., 2016; Moosavi-Dezfooli et al., 2016) as well as defenses against these (see (Schott et al., 2018) for an overview). Apart from the application-focused viewpoint, the observed adversarial vulnerability offers non-obvious insights into the inner workings of neural networks. One particular method of defense is *adversarial training* (Madry et al., 2018), which aims to minimize a modified training objective. While this method – like all known approaches of defense – decreases the accuracy of the classifier, it is also successful in increasing the robustness to adversarial attacks, i.e. the perturbations need to be larger on average in order to change the classification decision.

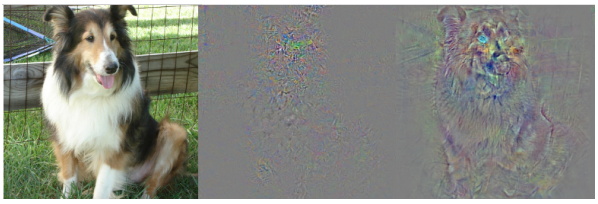

Figure 1. An image of a dog (left), the saliency maps of a highly non-adversarially-robust neural network (middle) and of a more robust network (right). We observe that the robust network gives a much clearer indication of what the classifier deems to be discriminative features. Details about saliency and the robustification are given in section 4. Most figures are best viewed on a screen.

(Tsipras et al., 2019) also notice that networks that are robustified in this way show interesting phenomena, which so far could not be explained. Neural networks usually exhibit very unstructured *saliency maps* (gradients of a classifier score with respect to the network's input (Simonyan et al., 2013)) which barely relate to the input image. On the other hand, saliency maps of robustified classifiers tend to be far more interpretable, in that structures in the input image also emerge in the corresponding saliency map, as exemplified in Figure 1. (Tsipras et al., 2019) describe this as an 'unexpected benefit' of adversarial robustness. In order to obtain a semantically meaningful visualization of the network's classification decision in non-robustified networks, the saliency map has to be aggregated over many different points in the vicinity of the input image. This can be achieved either via averaging saliency maps of noisy versions of the image (Smilkov et al., 2017) or by integrating along a path (Sundararajan et al., 2017). Other approaches typically employ modified backpropagation schemes in order to highlight the discriminative portions of the image. Examples of this include *guided backpropagation* (Springenberg et al., 2015)

[*]Equal contribution [1]Center for Industrial Mathematics, University of Bremen, Bremen, Germany [2]Work done at DAMTP, Cambridge. [3]DAMTP, University of Cambridge, Cambridge, United Kingdom. Correspondence to: Christian Etmann <cetmann@math.uni-bremen.de>, Sebastian Lunz <lunz@math.cam.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

and *deep Taylor decomposition* (Montavon et al., 2017).
In this paper, we show that the interpretability of the saliency maps of a robustified neural network is not only a side-effect of adversarial training, but a general property enjoyed by networks with a high degree of robustness to adversarial perturbations. We first demonstrate this principle for the case of a linear, binary classifier and show that the 'interpretability' is due to the image vector and the respective image gradient aligning. For the more general, non-linear case we empirically show that while this relationship is true on average, the linear theory and the non-linear reality do not always agree. We empirically demonstrate that the more linear the model is, the stronger the connection between robustness and alignment becomes.

2. Adversarial Robustness and Saliency Maps

Since adversarial perturbations are small perturbations that change the predicted class of a neural network, it makes sense to define the robustness towards adversarial perturbations via the distance of the unperturbed image to its nearest perturbed image, such that the classification is changed.

Definition 1. *Let $F : X \to C$ (with C finite) be a classifier over the normed vector space $(X, \|\cdot\|)$. We call*

$$\rho(x) = \inf_{e \in X} \{\|e\| : F(x+e) \neq F(x)\} \tag{1}$$

the (adversarial) robustness of F in the point x. *We call* $\mathbb{E}_{x\sim\mathcal{D}}[\rho(x)]$ *the* (adversarial) robustness of F over the distribution $\mathcal{D}$.

Put differently, the robustness of a classifier in a point is nothing but the distance to its closest decision boundary. Margin classifiers like support vector machines (Cortes & Vapnik, 1995) seek to keep this distance large for the training set, usually in order to avoid overfitting. (Sokolić et al., 2017) and (Elsayed et al., 2018) also apply this principle to neural networks via regularization schemes. We point out that our definition of adversarial robustness does not depend on the ground truth class label and – given feasible computability – can approximately be calculated even on unlabelled data.
In the following, we will always assume X to be a real, finite-dimensional vector space with the Euclidean norm. The proofs for the following theoretical statements are found in the appendix.

2.1. A Motivating Toy Example

We consider the toy case of a linear binary classifier $F(x) = \operatorname{sgn}(\Psi_z(x))$ with the so-called score function $\Psi_z(x) = \langle x, z\rangle$ and fixed $z \neq 0$, where $\langle\cdot,\cdot\rangle$ denotes the standard inner product on $\mathbb{R}^m$. A straightforward calculation (see appendix) shows that the adversarial robustness of F is given by

$$\rho(x) = \frac{|\langle x, z\rangle|}{\|z\|} = \frac{|\langle x, \nabla\Psi_z(x)\rangle|}{\|\nabla\Psi_z(x)\|}. \tag{2}$$

Unless stated otherwise, we will always denote with ∇ the gradient with respect to x. Note that $\rho(x) = \|x\| \cdot |\cos(\delta)|$, where δ is the angle between the vectors x and $\nabla\Psi_z(x)$. This implies that $\rho(x)$ grows with the alignment of x and z and is maximized if and only if x and z are collinear.
This motivates the following definition.

Definition 2 (Alignment)**.** *Let the binary classifier*

$$F : X \to \{-1, 1\}$$

be defined a.e. by $F(x) = \operatorname{sgn}(\Psi(x))$, where $\Psi : X \to \mathbb{R}$ is differentiable in x. We then call $\nabla\Psi$ the saliency map of F with respect to Ψ in x *and*

$$\alpha(x) := \frac{|\langle x, \nabla\Psi(x)\rangle|}{\|\nabla\Psi(x)\|}, \tag{3}$$

the alignment with respect to Ψ in x.

The alignment is a measure of how similar the input image x and the saliency map $\nabla\Psi(x)$ are. If $\|x\| = 1$, and x and $\nabla\Psi(x)$ are zero-centered, this coincides with the absolute value of their Pearson correlation. For a linear binary classifier, the alignment trivially increases with the robustness of the classifier.

Generalizing from the linear to the affine case leads to a classifier of the form $F(x) = \operatorname{sgn}(\langle x, z\rangle + b)$, whose robustness in x is

$$\rho(x) = \frac{|\langle x, z\rangle + b|}{\|z\|}.$$

In this case the robustness and alignment do not coincide anymore. In order to connect these two diverging concepts, we offer two alternative viewpoints. On the one hand, we can trivially bound the robustness via the triangle inequality

$$\rho(x) \leq \alpha(x) + \frac{|b|}{\|z\|}. \tag{4}$$

This is particularly meaningful if $|b|/\|z\|$ is small in comparison to $\alpha(x)$. Alternatively, one can connect the robustness to the alignment at a different point $\xi = x + \frac{b}{\|z\|}\frac{z}{\|z\|}$, leading to the relation

$$\rho(x) = \alpha(\xi). \tag{5}$$

In the affine case this approach simply amounts to a shift of the data that is uniform over all data points x. We will see how these two viewpoints lead to different bounds in the non-linear case later.

2.2. The General Case

We now consider the general, n-class case.

Definition 3 (Alignment, Multi-Class Case)**.** *Let*

$$\Psi = (\Psi^1, \dots, \Psi^n) : X \to \mathbb{R}^n$$

be differentiable in x. Then for an n-class classifier defined a.e. by

$$F(x) = \arg\max_i \Psi^i(x), \tag{6}$$

we call $\nabla\Psi^{F(x)}$ the saliency map of F. *We further call*

$$\alpha(x) := \frac{|\langle x, \nabla\Psi^{F(x)}(x)\rangle|}{\|\nabla\Psi^{F(x)}(x)\|}, \tag{7}$$

the alignment with respect to Ψ in x.

2.2.1. Linearized Robustness

In general the distance to the decision boundary $\rho(x)$ can be unfeasible to compute. However, for classifiers built on locally affine score functions – such as most neural networks using ReLU or leaky ReLU activations – $\rho(x)$ can easily be computed, provided the locally affine region is sufficiently large. To quantify this, define the radius of the locally affine component of F around x as

$$l(x) = \sup\{r \mid \forall i : \Psi^i \text{ affine in } B_r(x)\},$$

where $B_r(x)$ is the open ball of radius r around x with respect to the Euclidean metric.

Lemma 1. *Let F be a classifier with locally affine score function Ψ. Assume $l(x) \geq \rho(x)$. Then*

$$\rho(x) = \min_{j \neq i^*} \frac{\Psi^{i^*}(x) - \Psi^j(x)}{\|\nabla\Psi^{i^*}(x) - \nabla\Psi^j(x)\|}, \tag{8}$$

for $i^* := F(x)$ the predicted class at x.

Similar identities were previously also independently derived in (Elsayed et al., 2018) and (Jakubovitz & Giryes, 2018).

Note that while nearly all state-of-the art classification networks are piecewise affine, the condition $l(x) \geq \rho(x)$ is typically violated in practice. However, the lemma can still hold *approximately* as long as the linear approximation to the network's score functions is sufficiently good in the relevant neighbourhood of x. This motivates the definition of the *linearized (adversarial) robustness* $\tilde{\rho}$.

Definition 4 (Linearized Robustness)**.** *Let $\Psi(x)$ be the differentiable score vector for the classifier F in x. We call*

$$\tilde{\rho}(x) := \min_{j \neq i^*} \frac{\Psi^{i^*}(x) - \Psi^j(x)}{\|\nabla\Psi^{i^*}(x) - \nabla\Psi^j(x)\|}, \tag{9}$$

the linearized robustness *in x, where $i^* := F(x)$ is the predicted class at point x.*

We later show that the two notions lead to very similar results, even if the condition $l(x) \geq \rho(x)$ is violated.

2.2.2. Reducing the Multi-Class Case

In this section, we introduce a toolset which helps bridge the gap between the alignment and the linearized robustness of a multi-class classifier. In the following, for fixed x, let $i^* := F(x)$ and j^* be the minimizer in (9). We can assign F in x a *binarized classifier* $F_x^\dagger$ with

$$F_x^\dagger(y) := \operatorname{sgn}\left(\Psi_x^\dagger(y)\right), \tag{10}$$

where $\Psi_x^\dagger(y) := \Psi^{i^*}(y) - \Psi^{j^*}(y)$. Its linearized robustness in $y = x$ is the same as for F. The *binarized saliency map*, $\nabla\Psi_x^\dagger(x) = \nabla_y\Psi_x^\dagger(y)|_{y=x}$ and the respective alignment,

$$\alpha^\dagger(x) = \frac{|\langle x, \nabla(\Psi^{i^*} - \Psi^{j^*})(x)\rangle|}{\|\nabla(\Psi^{i^*} - \Psi^{j^*})(x)\|}, \tag{11}$$

which we call *binarized alignment*, offer an alternative, natural perspective of the above considerations. This is because for classifiers as defined in (6), the actual score values do not necessarily carry any information about the classification decision, whereas the score differences do. While, roughly speaking, $\nabla\Psi^{i^*}$ tells us what F 'thinks' makes x a member of its predicted class, $\nabla\Psi_x^\dagger(x)$ carries information what sets x apart from its closest neighboring class (according to linearization).

In the special case of a linear, multi-class classifier, we have

$$\rho(x) = \tilde{\rho}(x) = \alpha^\dagger(x)$$

and in the linear, binary case $\Psi(x) = (\langle x, z\rangle, -\langle x, z\rangle)$, even

$$\alpha(x) = \alpha^\dagger(x).$$

3. Decompositions and Bounds for Neural Networks

3.1. Homogeneous Decomposition

In the previous chapter we have seen that in the case of binary classifiers, the robustness and binarized alignment coincide for linear score functions. However, requiring Ψ to be linear is a stronger assumption than necessary to deduce the result: It is in fact sufficient for Ψ to be *positive one-homogeneous*. Any such function satisfies $\Psi(ax) = a\Psi(x)$ for all $a > 0$ and x.

Lemma 2 (Linearized Robustness of Homogeneous Classifiers)**.** *Consider a classifier F with positive one-homogeneous score functions. Then*

$$\tilde{\rho}(x) = \alpha^\dagger(x). \tag{12}$$

In particular, most feedforward neural networks with (leaky) ReLU activations *without biases* are positive one-homogeneous. This observation motivates to split up any classifier built on neural networks into a homogeneous term and the corresponding remainder, leading to the following decomposition result.

Theorem 1 (Homogeneous Decomposition of Neural Networks)**.** *Let $\Psi^i_{\Theta,b}$ be any logit of a neural network with ReLU activations (of class $\mathcal{N}$ in the appendix). Denote by Θ the linear filters and by b the bias terms of the network. Then*

$$\begin{aligned}\Psi^i_{\Theta,b}(x) &= \langle x, \nabla_x \Psi^i_{\Theta,b}(x)\rangle + \langle b, \nabla_b \Psi^i_{\Theta,b}(x)\rangle \\ &= \langle x, \nabla_x \Psi^i_{\Theta,b}(x)\rangle + \sum_k b_k \partial_{b_k} \Psi^i_{\Theta,b}(x). \end{aligned} \tag{13}$$

Note that the above vector b includes the running averages of the means for batch normalization. For ReLU networks, the remainder term $\beta^i(x) := \langle b, \nabla_b \Psi^i_{\Theta,b}(x)\rangle$ is locally constant, because it changes only when x enters another locally linear region. For ease of notation, we will now drop the subscripts Θ and b.

3.2. Pointwise Bounds

In section 2.1, we introduced two different viewpoints for affine linear, binary classifiers which connect the robustness to the alignment. In a similar vein to inequality (4) and equality (5), upper bounds to the linearized robustness depending on the alignment can be given for neural networks. In the following, we will write $\overline{v} := v/\|v\|$ for $v \neq 0$. Again, in the following we fix x and write $i^* = F(x)$ and j^* for the minimizer in j from equation (9).

Theorem 2. *Let $g := \nabla \Psi^{i^*}(x)$. Furthermore, let $g^\dagger := \nabla(\Psi^{i^*} - \Psi^{j^*})(x)$ and $\beta^\dagger := \beta^{i^*}(x) - \beta^{j^*}(x)$. Then*

$$\tilde{\rho}(x) \leq \alpha^\dagger(x) + \frac{|\beta^\dagger|}{\|g^\dagger\|} \tag{14}$$

$$\leq \alpha(x) + \|x\| \cdot \|\overline{g}^\dagger - \overline{g}\| + \frac{|\beta^\dagger|}{\|g^\dagger\|}. \tag{15}$$

Distances on the unit sphere (such as $\|\overline{g}^\dagger - \overline{g}\|$) can be converted to angles through the law of cosines. For the above inequalities to be reasonably tight, the angle between g and $g^\dagger$ needs to be small and $|\beta^\dagger|/\|g^\dagger\|$ needs to be small in comparison to $\alpha^\dagger(x)$. In this case, the alignment should roughly increase with the linearized robustness.

Theorem 3. *Let $\xi := x + \frac{\beta^\dagger}{\|g^\dagger\|}\frac{g^\dagger}{\|g^\dagger\|}$ and $\gamma := \nabla \Psi^{i^*}(\xi)$, with $g^\dagger$ and $\beta^\dagger$ defined as in the previous theorem. Then*

$$\tilde{\rho}(x) \leq \frac{|\langle \xi, \gamma\rangle|}{\|\gamma\|} + \|\xi\| \cdot \|\overline{g}^\dagger - \overline{\gamma}\|, \tag{16}$$

and if additionally $F(x) = F(\xi)$, then

$$\tilde{\rho}(x) \leq \alpha(\xi) + \|\xi\| \cdot \|\overline{g}^\dagger - \overline{\gamma}\|.$$

Depending on the sign of $\beta^\dagger$, the shifted image ξ can either be understood as a gradient ascent or descent iterate for maximizing/minimizing $\Psi^{i^*} - \Psi^{j^*}$. This theorem assimilates $\beta^\dagger(x)$ into x, providing an upper bound to $\tilde{\rho}(x)$ that depends on $\alpha(\xi)$. The sensibility of this hinges on ξ being reasonably close to x and γ having a low angle with $g^\dagger$.

If the error terms in inequalities (14), (15) and (16) are small, these inequalities thus provide a simple illustration why more robust networks yield more interpretable saliency maps.

Nevertheless, the right-hand side may be much larger than $\tilde{\rho}(x)$, if the inner product between an image and its respective saliency map are almost orthogonal. This is because the Cauchy-Schwarz inequality (see the proofs in the appendix) provides a large upper bound in this case. The inequalities rather serve as an explanation of how the various terms of alignment may deviate from the linearized robustness in the case of a neural network.

3.3. Alignment and Interpretability

The above considerations demonstrate how an increase in robustness may induce an increase in the alignment between an input image and its respective saliency map. The initial observation – which was previously described as an increase in *interpretability* – may thus be ascribed to this phenomenon. This is especially true in the case of natural images, as exemplified in Figure 1. There, what a human observer would deem an increase in interpretability, expresses itself as discriminative portions of the original image reappearing in the saliency map, which naturally implies a stronger alignment. The concepts of alignment and interpretability should however not be conflated completely: In the case of quasi-binary image data like MNIST, 0-regions of the image render the inner product in equation (7) invariant with respect to the saliency map in this region, even if the saliency map e.g. assigns relevance to the absence of a feature in this region. Note however that the saliency map in this region still influences the alignment term through the division by its norm. Additionally, the alignment is also not invariant to the images' representation (color space, shifts, normalization etc.). Still, for most types of image data an increase in alignment in discriminative regions should coincide with an increase in interpretability.

4. Experiments

In order to validate our hypothesis, we trained several models of different adversarial robustness on both MNIST (LeCun et al., 1990) and ImageNet (Deng et al., 2009) using double backpropagation (Drucker & Le Cun, 1992). For a neural network f_θ with a softmax output layer, this amounts

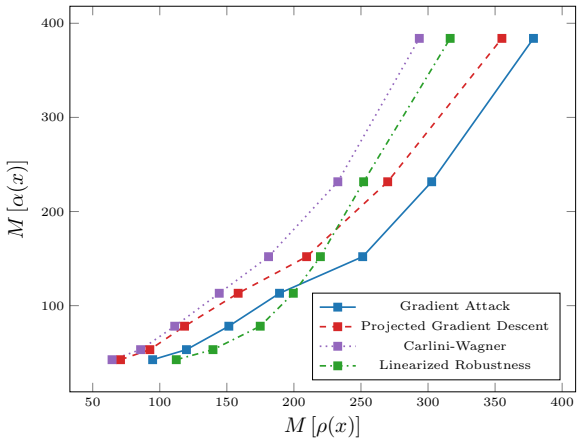

Figure 2. The median alignment increases with the median robustness of the model on ImageNet. Furthermore, the more elaborate attacks consistently find smaller adversarial perturbations than the simple gradient attack. The linearized robustness estimator provides a rather realistic estimation of the algorithmically calculated robustness.

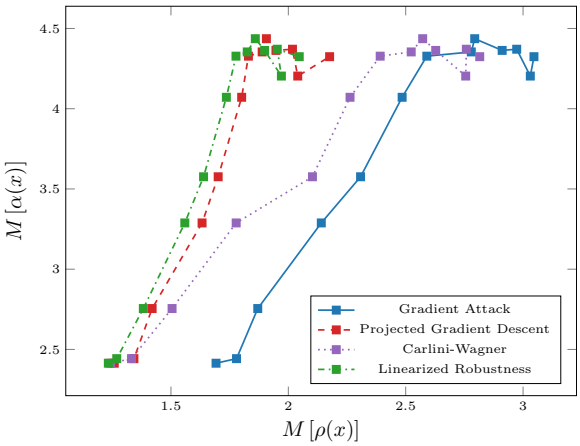

Figure 3. Similar to Figure 2, the median alignment increases with the median robustness of the model on MNIST. Towards the end, some saturation effects are visible.

to minimizing the modified loss

$$\frac{1}{N}\sum_{i=1}^{N}\left[\mathcal{L}(f_\theta(x^{(i)}),y^{(i)})+\lambda\cdot\|\nabla\,\mathcal{L}(f_\theta(x^{(i)}),y^{(i)})\|^2\right] \tag{17}$$

over the parameters $\theta=(\Theta,b)$. Here, $\{(x^{(i)},y^{(i)})\}_{i=1,\dots,N}$ is the training set and $\mathcal{L}$ denotes the negative log-likelihood error function. The hyperparameter $\lambda\geq 0$ determines the strength of the regularization. Note that this penalizes the local Lipschitz constant of the loss. As (Simon-Gabriel et al., 2018) demonstrate, double backpropagation makes neural networks more resilient to adversarial attacks. By varying λ, we can easily create models of different adversarial robustness for the same dataset, whose properties we can then compare. (Anil et al., 2018) previously noted that Lipschitz constrained networks exhibit interpretable saliency maps (without an explanation), which can be regarded as a side-effect of the increase in adversarial robustness.

For the MNIST experiments, we trained each of our 16 models on an NVIDIA 1080Ti GPU with a batch size of 100 for 200 epochs, covering the regularization hyperparameter range from 10 to 180,000, before the models start to degenerate. The used architecture is found in the appendix.

For the experiments on ImageNet, we fine-tuned the pre-trained ResNet50 model from (He et al., 2016) over 35 epochs on 2 NVIDIA P100 GPUs with a total batch size of 32. We used stochastic gradient descent with a learning rate of 0.0001 and momentum of 0.99. The learning rate was divided by 10 whenever the error stopped improving. For the regularization parameter, we chose $\lambda=10^4,10^{4.5},\dots,10^7$. The experiments were implemented in Tensorflow (Abadi et al., 2015).

4.1. Robustness and Alignment

For checking the relation between the alignment and robustness of a neural network, we created 1000 adversarial examples per model on the respective validation set. This was realized using the python library *Foolbox* (Rauber et al., 2017), which offers pre-defined adversarial attacks, three of which we used in this paper: The `GradientAttack` performs a line search for the closest adversarial example along the direction of the loss gradient. `L2BasicInterativeAttack` implements the projected gradient descent attack from (Kurakin et al., 2016) for the Euclidean metric. Similarly, `CarliniWagnerL2Attack` (*CW-attack*) is the attack introduced in (Carlini & Wagner, 2017) suited for finding the closest adversarial example in Euclidean metric. Additionally, we calculated the linearized robustness $\tilde{\rho}(x)$, which entails calculating n gradients per image for an n-class problem.

In Figures 2 and 3, we investigate how the median alignment depends on the medians of the different conceptions of robustness. We opted in favor of the median (M) instead of the arithmetic mean due to its increased robustness to outliers, which occurred especially when using the gradient attack. In the case of ImageNet (Figure 2), an increase in median alignment with the median robustness is clearly visible for all three estimates of the robustness. On the other hand, the alignment for the MNIST data increases with the robustness as well, but seems to saturate at some point. We will offer an explanation for this phenomenon later.

We now consider the pointwise connection between robustness and alignment. In Figure 4 the two variables are highly-correlated for a model trained on MNIST, pointing towards the fact that the network behaves very similarly to a positive one-homogeneous function. There is however no

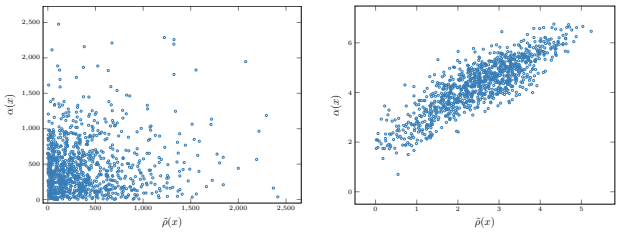

Figure 4. The pointwise relationship between $\tilde{\rho}(x)$ and $\alpha(x)$, exemplified on a model trained on ImageNet (left) and MNIST (right). While the two properties are well-correlated on MNIST (fitting the 'averaged' view from Figure 3), there is no visible correlation in the case of ImageNet.

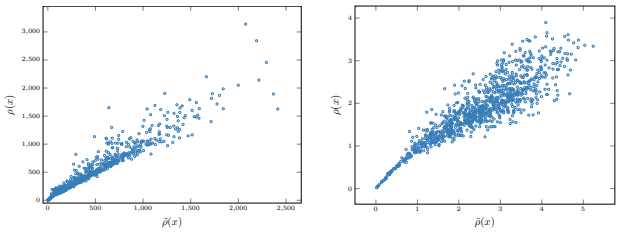

Figure 5. The pointwise relationship between $\tilde{\rho}(x)$ and $\rho(x)$, each calculated for 1000 validation points on a model trained on ImageNet (left) and MNIST (right). $\rho(x)$ was approximately calculated using the CW-attack. In both cases, the correlation is high.

visible correlation between them on the ImageNet model, which is a consistent behavior throughout the whole experiment cohort. We will later analyse the source of this behavior. The increase in median alignment for ImageNet, $M\left[\alpha(x)\right] = M\left[|\langle x, \overline{g}\rangle|\right]$, can still be explained by a statistical argument: If $M\left[\langle x, \overline{g}\rangle\right] = 0$, as approximately true in our ImageNet model, then $M\left[\alpha(x)\right]$ is the median absolute deviation of $\langle x, \overline{g}\rangle$. In other words, the graph for ImageNet in Figure 4 depicts the dispersion of $\langle x, \overline{g}\rangle$. The above observations also hold well for the binarized alignment.

In Figure 5 a tight correlation between $\tilde{\rho}(x)$ and $\rho(x)$ becomes evident. Here, the latter has been calculated using the CW-attack. The linearized robustness model $\tilde{\rho}$ is hence an adequate approximation of the actual robustness ρ, even for the highly non-linear neural network models used on ImageNet. Finally note that all used attacks lead to the same general behavior of all quantities investigated (see Figures 2 and 3).

4.2. Explaining the Observations

In the last section, we observed some commonalities between the experiments on ImageNet and MNIST, but also some very different behaviors. In particular, two aspects stand out: Why does the median alignment steadily increase for the observed ImageNet experiments, whereas on MNIST this stagnates at some point (Figures 2 and 3)? Furthermore, why are $\tilde{\rho}(x)$ and $\alpha(x)$ so highly-correlated on MNIST but almost uncorrelated on ImageNet (Figure 4)? We turn to Theorems 2 and 3 for answers.
Theorem 2 states that

$$\tilde{\rho}(x) \leq \alpha^{\dagger}(x) + \frac{|\beta^{\dagger}|}{\|g^{\dagger}\|}, \tag{18}$$

where $\beta^{\dagger}$ is the locally constant term and $g^{\dagger}$ is the saliency map of the binarized classifier and $\overline{v} = v/\|v\|$ for $v \neq 0$. In Figure 6, we check how strongly the right-hand side of inequality (18) is dominated by $\alpha^{\dagger}(x)$, i.e. how large the influence of the locally linear term is in comparison to the locally constant term. For ImageNet, this ratio increases from below 0.55 to almost 0.85, pointing towards a model increasingly governed by its linearized part. On MNIST, this ratio strongly decreases over the robustness's range. Note however that in the weakly regularized MNIST models, the right hand side is extremely dominated by the median alignment in the first place.

A similar analysis can be performed for the second inequality from Theorem 2,

$$\tilde{\rho}(x) \leq \alpha(x) + \|x\| \cdot \|\overline{g}^{\dagger} - \overline{g}\| + \frac{|\beta^{\dagger}|}{\|g^{\dagger}\|}, \tag{19}$$

which additionally makes a step from binarized alignment to (conventional) alignment.

This leads to an additional error term, making the bound significantly less tight than in the previous case. In particular, the proportion of the alignment α on the right-hand side diminishes, confirming our prediction from section 3.2. Nevertheless, the qualitative behaviors is similar to the previous case, with the $\alpha(x)$ taking up an increasing fraction of the right-hand with increasing robustness. For MNIST data, the ratio varies little compared to the ratio from the last inequality. This indicates that the remainder term $\|\overline{g}^{\dagger} - \overline{g}\|$ does not change too strongly over the set of MNIST experiments compared to $\alpha(x)$. We thus deduce that the qualitative relationship between robustness and alignment is fully governed by the error term introduced in (18), i.e. the locally constant term of the logit.

We now do the same for the inequality in Theorem 3, which states that

$$\tilde{\rho}(x) \leq \frac{|\langle \xi, \gamma\rangle|}{\|\gamma\|} + \|\xi\| \cdot \|\overline{g}^{\dagger} - \overline{\gamma}\| \tag{20}$$

for $\xi = x + \frac{b^{\dagger}}{\|g^{\dagger}\|}\frac{g^{\dagger}}{\|g^{\dagger}\|}$ and $\gamma = \nabla\Psi^{F(x)}(\xi)$, which gets rid of the additive term $|\beta^{\dagger}|/\|g^{\dagger}\|$ from (18). Again, in the case of ImageNet $|\langle \xi, \overline{\gamma}\rangle|$ grows more quickly in comparison to $\|\overline{g}^{\dagger} - \overline{\gamma}\|$, the distance of the normalized gradients, whereas

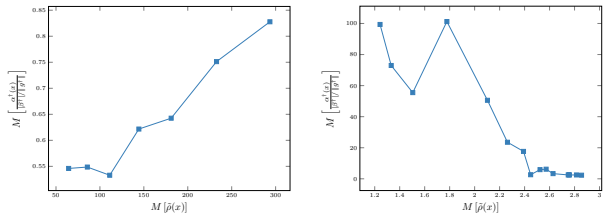

Figure 6. Comparing the size of the summands of inequality (18) for the various experiments. In the case of ImageNet (left), $\alpha^\dagger(x)$ takes up an increasing fraction of the right-hand side of the inequality. For MNIST (right), this portion tends to strongly decrease with the robustness. Note however that in this case, $\alpha^\dagger(x)$ starts out vastly dominating the right-hand side.

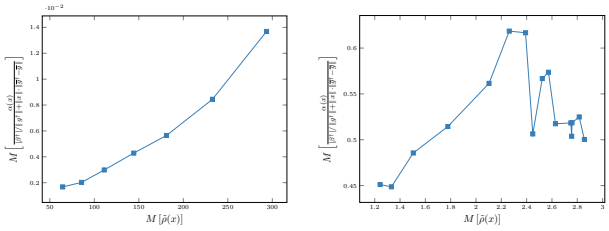

Figure 7. Comparing the size of the summands of inequality (19) for the various experiments. For the ImageNet experiments (left), the portion of $\alpha(x)$ of the right-hand side of the inequality increases roughly 7-fold. For MNIST (right), this portion stays roughly constant compared to the variation from Figure 6.

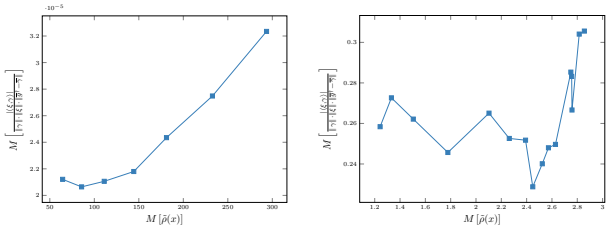

Figure 8. Comparing the size of the summands of inequality (20) for the various experiments. In the case of ImageNet (left), the alignment in ξ takes up an increasingly large portion of the right-hand side of the inequality. For MNIST (right), this portion stays roughly constant.

their ratio is approximately constant for MNIST data.

To conclude, we have seen that the upper bounds from Theorems 2 and 3 provide valuable information in which ways both the experiments on ImageNet and MNIST are influenced by the respective terms. In the case of ImageNet, we consistently see the alignment terms growing more quickly than the other terms. This might indicate that the growth in alignment stems not only from the growth in the robustness alone, but also from the model becoming increasingly similar to our idealized toy example. In other words, not only does the robustness make the alignment grow, but the connection between these two properties becomes stronger in the case of ImageNet. This is in agreement with the seemingly superlinear growth of the median alignment in Figure 2.

It is not surprising that a classifier for a problem as complex as ImageNet is highly non-linear, which makes the (pointwise) connection between alignment and robustness rather loose. We hence conjecture that the imposed regularization increasingly restricts the models to be more linear, thereby making them more similar to our initial toy example.

For MNIST, the regularization seems to have the opposite effect: As seen in Figure 6, the binarized alignment initially dwarfs the correction term $|\beta^\dagger|/\|g^\dagger\|$ introduced by the locally constant portion of the binarized logit $\Psi_x^\dagger(x)$. As the network becomes more robust, $\Psi_x^\dagger(x)$ is apparently not dominated by the linear terms anymore, while the influence of the locally constant terms (i.e. $\beta^\dagger$) increases. This hypothesis seems sensible, considering MNIST is a very simple problem which we tackled with a comparatively shallow network. This can be expected to yield a model with a low degree of non-linearity. The penalization of the local Lipschitz constant here seems to have the effect of requiring larger locally constant terms $|\beta^\dagger|$, in contrast to the models trained on ImageNet.

We check the validity of these claims by tracking the median size of $|\langle x, g^\dagger\rangle|$ against the median size of $|\Psi_x^\dagger(x)|$ in Figure 9. On MNIST, $M\left[|\langle x, g^\dagger\rangle|\right]$ starts out at approximately 40% of $M\left[|\Psi_x^\dagger(x)|\right]$ and at the end rises to almost 100%. Note that this does not indicate that β^i is typically close to 0 for all i, just that $\beta^\dagger$ is, compared to $\langle x, g^\dagger\rangle$.

On MNIST, this ratio is close to 1 up until $M\left[\tilde{\rho}(x)\right] \approx 2.4$, when it suddenly and quickly falls below 0.5. This drop is consistent with what we see in Figure 3: At around the same point this drop occurs, the alignment starts to saturate. While an increase in the model's median robustness should imply an increase in the model's median alignment, the deviation from linearity weakens the connection between robustness and alignment, such that the two effects roughly cancel out.

In Figure 10, we provide examples for the different gradient

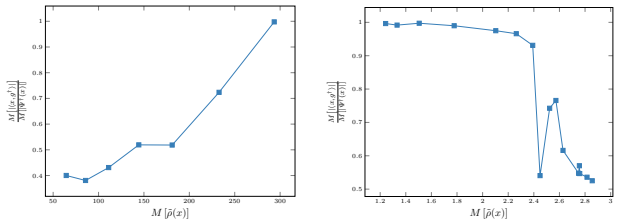

Figure 9. On the ImageNet experiments, the linear term $|\langle x, g^\dagger \rangle|$ takes up an increasing portion of the binarized score $\Psi^\dagger(x)$. In the case of MNIST, $\Psi^\dagger(x)$ is completely dominated by the linear term, before its influence decreases sharply at $M\,[\tilde{\rho}(x)] \approx 2.4$.

concepts we introduced in Theorems 2 and 3, both for the most robust and non-robust network from our experiment cohort.

5. Conclusion and Outlook

In this paper, we investigated the connection between a neural network's robustness to adversarial attacks and the interpretability of the resulting saliency maps. Motivated by the binary, linear case, we defined the alignment α as a measure of how much a saliency map matches its respective image. We hypothesized that the perceived increase in interpretability is due to a higher alignment and tested this hypothesis on models trained on MNIST and ImageNet. While on average, the proposed relation holds well, the connection is much less pronounced for individual points, especially on ImageNet. Using some upper bounds for the robustness of a neural network, which we derived using a decomposition theorem, we arrived at the conclusion that the strength of this connection is strongly linked with how similar to a linear model the neural network is locally. As ImageNet is a comparatively complex problem, any sufficiently accurate model is bound to be very non-linear, which explains the difference to MNIST.

While this paper shows the general link between robustness and alignment, there are still some open questions. Since we only used one specific robustification method, further experiments should determine the influence of this method. One could explore, whether a different choice of norm leads to different observations. Another future direction of research could be to investigate the degree of (non-)linearity and its connection to this topic. While Theorems 2 and 3 illustrate how the pointwise linearized robustness and alignment may diverge, depending on terms like g, $g^\dagger$, γ and $\beta^\dagger$, a more in-depth look should focus on *why* and *when* these terms have a certain relationship to each other.

From a methodological standpoint, the discovered connection may also serve as an inspiration for new adversarial defenses, where not only the robustness but also the alignment is taken into account. One way of increasing the alignment directly would be through the penalty term

$$\lambda \left(\|x\|^2 \|\nabla \Psi^i(x)\|^2 - \langle x, \nabla \Psi^i(x) \rangle^2 \right),$$

which is bounded from below by 0 via the Cauchy-Schwarz inequality. Any robustifying effects of the increased alignment may however be confounded with the Lipschitz-penalty that the first summand effectively introduces, which necessitates a careful experimental evaluation.

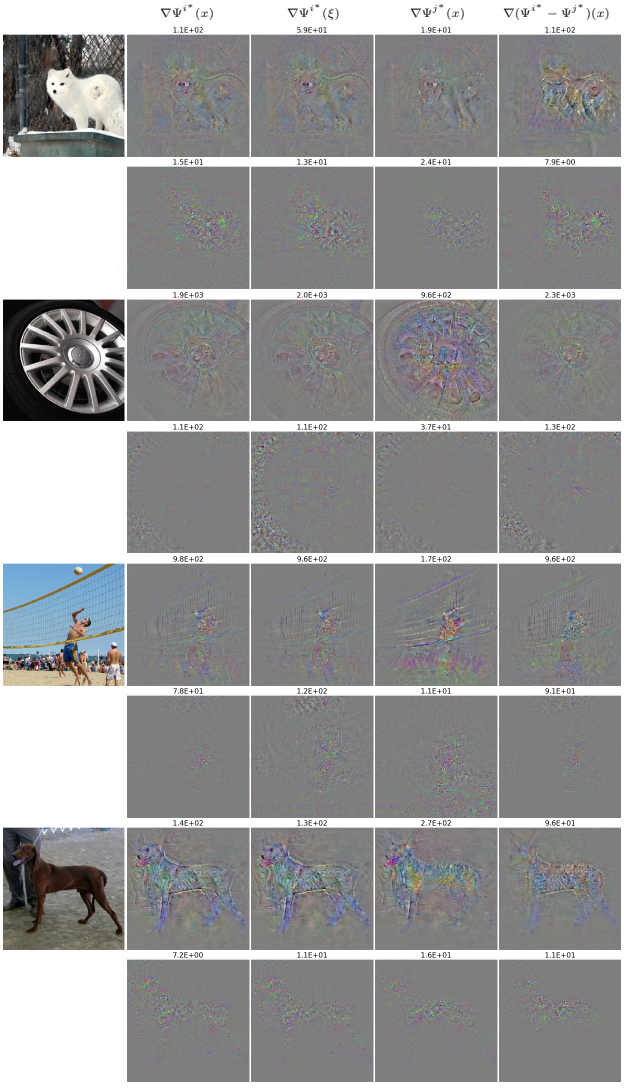

Figure 10. Selected examples from the ImageNet validation set of the different gradients and their respective alignments with x, respectively ξ. The odd rows are generated with the most robust ImageNet classifier, whereas the even rows are generated by the least robust classifier. The gradient images are individually scaled to fit the color range $[0, 255]$.

Acknowledgements

CE and PM acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 281474342: 'RTG π^3 - Parameter Identification - Analysis, Algorithms, Applications'. The work by SL was supported by the EPSRC grant EP/L016516/1 for the University of Cambridge Centre for Doctoral Training, the Cambridge Centre for Analysis and by the Cantab Capital Institute for the Mathematics of Information. CBS acknowledges support from the Leverhulme Trust projects on Breaking the non-convexity barrier and on Unveiling the Invisible, the Philip Leverhulme Prize, the EPSRC grant Nr. EP/M00483X/1, the EPSRC Centre Nr. EP/N014588/1, the European Union Horizon 2020 research and innovation programmes under the Marie Skodowska-Curie grant agreement No 777826 NoMADS and No 691070 CHiPS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Quadro P6000 and a Titan Xp GPUs used for this research.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Anil, C., Lucas, J., and Grosse, R. Sorting out lipschitz function approximation. *arXiv preprint arXiv:1811.05381*, 2018.

Carlini, N. and Wagner, D. A. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 39–57, 2017.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. Ieee, 2009.

Drucker, H. and Le Cun, Y. Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, 3(6):991–997, 1992.

Elsayed, G. F., Krishnan, D., Mobahi, H., Regan, K., and Bengio, S. Large margin deep networks for classification. 2018. URL https://arxiv.org/pdf/1803.05598.pdf.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv 1412.6572*, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jakubovitz, D. and Giryes, R. Improving dnn robustness to adversarial attacks using jacobian regularization. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pp. 396–404, 1990.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436, 2015.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. 2018.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

Rauber, J., Brendel, W., and Bethge, M. Foolbox v0. 8.0: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

Schott, L., Rauber, J., Brendel, W., and Bethge, M. Robust perception through analysis by synthesis. *CoRR*, abs/1805.09190, 2018. URL http://arxiv.org/abs/1805.09190.

Simon-Gabriel, C.-J., Ollivier, Y., Schölkopf, B., Bottou, L., and Lopez-Paz, D. Adversarial vulnerability of neural networks increases with input dimension. *arXiv preprint arXiv:1802.01421*, 2018.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Sokolić, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017.

Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. URL http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. 2014.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.