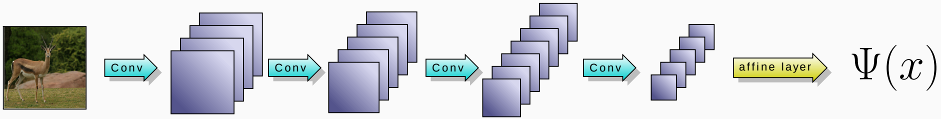
On the Connection Between Adversarial Robustness and Saliency Map Interpretability

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

13th June, 2019

1: ZeTeM, University of Bremen, 2: Cambridge Image Analysis, University of Cambridge, 3: Work done at Cambridge

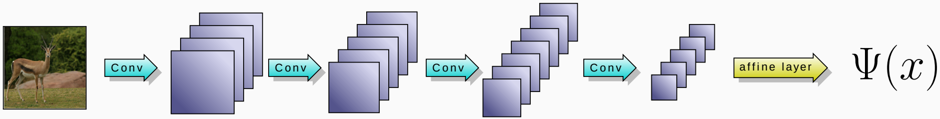
Saliency Maps



For a logit $\Psi^i(x)$, we call its gradient $r\,\Psi^i(x)$ the *saliency map* in x.

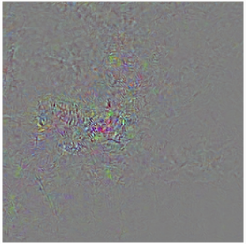
It *should* show us the discriminative portions of the image.

Saliency Maps



For a logit $\Psi^i(x)$, we call its gradient $r\,\Psi^i(x)$ the *saliency map* in x.
It *should* show us the discriminative portions of the image.

Original Image

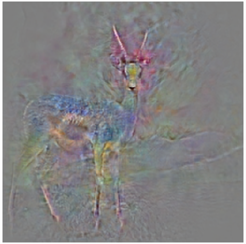


Saliency map of a ResNet50

An Unexplained Phenomenon

Models trained to be more robust to adversarial attacks seem to exhibit 'interpretable' saliency maps[1]

Original Image



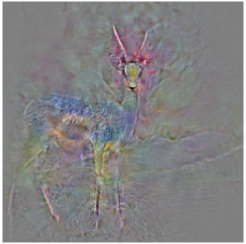
Saliency map of a robustified ResNet50

[1] Tsipras et al, 2019: 'Robustness may be at odds with accuracy.'

An Unexplained Phenomenon

Models trained to be more robust to adversarial attacks seem to exhibit 'interpretable' saliency maps[1]

Original Image



Saliency map of a robustified ResNet50

This phenomenon has a remarkably simple explanation!

[1] Tsipras et al, 2019: 'Robustness may be at odds with accuracy.'

Explaining the Interpretability Puzzle

We call

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

the *adversarial robustness* of the classifier F (with respect to euclidean norm $\|\cdot\|$).

- Adversarial attacks are tiny perturbations that 'fool' the classifier

Explaining the Interpretability Puzzle

We call

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

the *adversarial robustness* of the classifier F (with respect to euclidean norm $\| \cdot \|$).

- Adversarial attacks are tiny perturbations that 'fool' the classifier
- A higher robustness to these attacks $\Rightarrow$ greater distance to the decision boundary

Explaining the Interpretability Puzzle

We call

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

the *adversarial robustness* of the classifier F (with respect to euclidean norm $\|\cdot\|$).

- Adversarial attacks are tiny perturbations that 'fool' the classifier
- A higher robustness to these attacks $\Rightarrow$ greater distance to the decision boundary
- A larger distance to the decision boundary results in a lower angle between x and $\nabla\Psi^i(x)$

Explaining the Interpretability Puzzle

We call

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

the *adversarial robustness* of the classifier F (with respect to euclidean norm $\|\cdot\|$).

- Adversarial attacks are tiny perturbations that 'fool' the classifier
- A higher robustness to these attacks $\Rightarrow$ greater distance to the decision boundary
- A larger distance to the decision boundary results in a lower angle between x and $\nabla\Psi^i(x)$
- We perceive this as a higher visual alignment between image and saliency map

Explaining the Interpretability Puzzle

We call

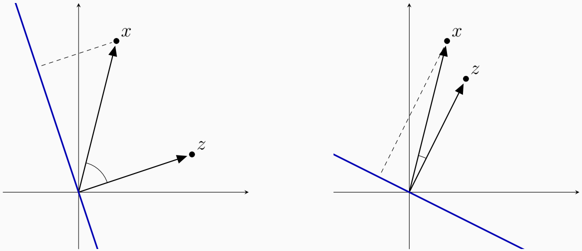
$$\rho(x) = \inf_{e \in X} \{\|e\| : F(x+e) \neq F(x)\}$$

the *adversarial robustness* of the classifier F (with respect to euclidean norm $\|\cdot\|$).

- Adversarial attacks are tiny perturbations that 'fool' the classifier
- A higher robustness to these attacks $\Rightarrow$ greater distance to the decision boundary
- A larger distance to the decision boundary results in a lower angle between x and $\nabla \Psi^i(x)$
- We perceive this as a higher visual alignment between image and saliency map

. . . but not quite

A Simple Toy Example



First, we consider a linear, binary classifier

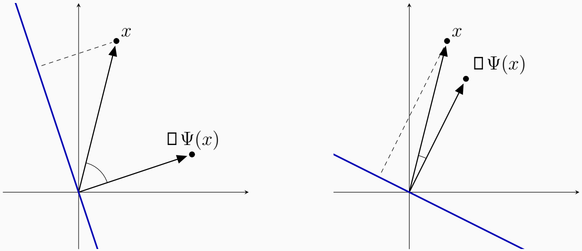
$$F(x) = \operatorname{sgn}\left(\Psi(x)\right),$$

where $\Psi(x) := \langle x, z \rangle$ for some z. Then

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

Note that $\rho(x) = \|x\| \, |\cos(\delta)|$, where δ is the angle between x and z.

A Simple Toy Example



First, we consider a linear, binary classifier

$$F(x) = \operatorname{sgn}\left(\Psi(x)\right),$$

where $\Psi(x) := \langle x, z \rangle$ for some z. Then

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

Note that $\rho(x) = \|x\| \cdot |\cos(\delta)|$, where δ is the angle between x and z.

Alignment

Definition (Alignment)

Let $\Psi = (\Psi^1, \ldots, \Psi^n) : X \to \mathbb{R}^n$ be differentiable in x. Then for an n-class classifier defined a.e. by $F(x) = \arg\max_i \Psi^i(x)$, we call $\nabla \Psi^{F(x)}$ the *saliency map of F*. We further call

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

the *alignment with respect to Ψ in x*.

For binary, linear models by construction: $\rho(x) = \alpha(x)$

Alignment

Definition (Alignment)

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

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

the *alignment with respect to Ψ in x*.

For binary, linear models by construction: $\rho(x) = \alpha(x)$

....but already wrong for affine models.

How about neural nets?

There is no closed expression for robustness. One idea is to **linearize.**

Definition (Linearized Robustness)

Let $\Psi(x)$ be the differentiable score vector for the classifier F in x. We call

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

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

Bridging the Gap Between Linearized Robustness and Alignment

Using

- a homogeneous decomposition theorem
- the 'binarization' of our classifier

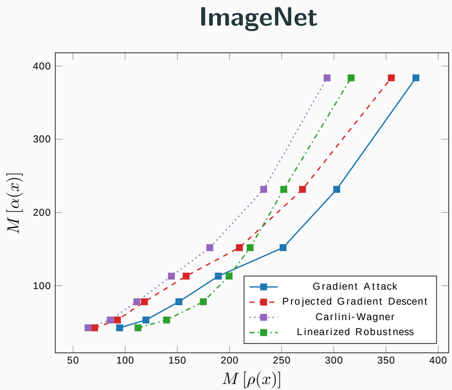
we get

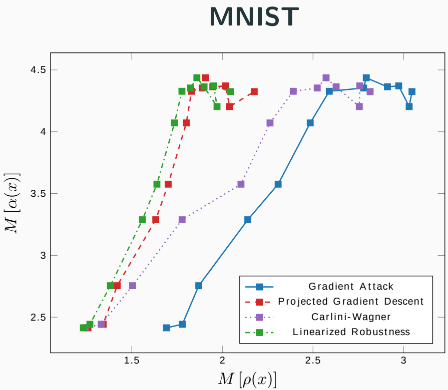
Theorem (Bound for general models)

Let $g := r\Psi^{i}\ (x)$. Furthermore, let $g^{y} := r\Psi^{y}_{x}(x)$ and β^{y} the non-homogeneous portion of Ψ^{y}_{x}. Denote by $\bar{v}$ the $k\ \ k$-normed $v\ 6\!\!= 0$. Then

$$\tilde{\rho}(x) \quad \alpha(x) + kxk \;\; k\overline{g}^{y} \quad \overline{g}k + \frac{j\beta^{y}j}{kg^{y}k}.$$

Experiments: Robustness vs. Alignment





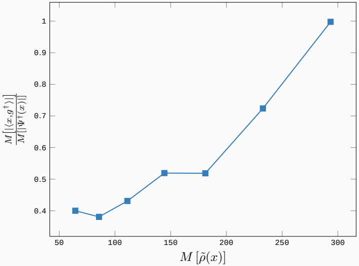
Linearized robustness is a reasonable approximation

Alignment increases with robustness

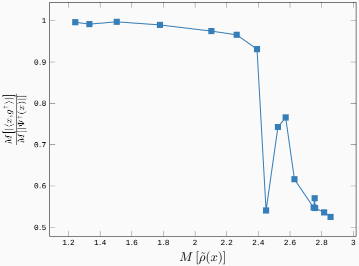
Superlinear growth for ImageNet and saturating effect on MNIST

Experiments: Explaining the Observations

ImageNet



MNIST



Fraction of homogeneous part of logit

The degree of homogeneity largely determines how strong the connection between α and $\tilde{\rho}$ is

ImageNet: higher robustness + more homogeneity = superlinear growth

MNIST: higher robustness + less homogeneity = effects start cancelling out

Thank you and see you at the poster!

Pacific Ballroom, #70