Verifying Attention Robustness of Deep Neural Networks against Semantic Perturbations

Satoshi Munakata
Fujitsu
Kanagawa, Japan
munakata.satosi@fujitsu.com

Caterina Urban
Inria & ENS | PSL & CNRS
Paris, France
caterina.urban@inria.fr

Haruki Yokoyama, Koji Yamamoto
and Kazuki Munakata
Fujitsu
Kanagawa, Japan

***Abstract*—In this paper, we propose the first verification method for *attention robustness*, i.e., the local robustness of the changes in the saliency-map against combinations of semantic perturbations. Specifically, our method determines the range of the perturbation parameters (e.g., the amount of brightness change) that maintains the difference between the actual saliency-map change and the expected saliency-map change below a given threshold value. Our method is based on linear activation region traversals, focusing on the outermost boundary of attention robustness for scalability on larger deep neural networks.**

I. Background

Classification Robustness. Deep neural networks (DNN) are dominant solutions in image classification; however, quality assurance is essential when DNNs are used in safety-critical systems [1]. Reference [2] reported that despite input images can be perturbed in the real world by various mechanisms, such as *brightness change* and *translation*; such *semantic perturbations* can unexpectedly change the classification labels for DNNs. Therefore, it is essential to verify *classification robustness (CR)* against semantic perturbations to be tolerable from an assurance point of view. Several methods have already been proposed to compute the range of perturbation parameters (e.g., the amount of brightness change and translation) that do not change classification labels [3], [4]. They defined the property of CR by the following predicate:

$$CR(x;\Theta) \overset{\text{def}}{=} \forall\theta \in \Theta . F\big(g(\theta, x)\big) = F(x) \tag{1}$$

where $x \in X \subseteq \mathbb{R}^{d_X}$ is an d_X-dimensional real vector of an input image, $\Theta \subset \mathbb{R}^{d_\Theta}$ is the range of d_Θ-dimensional perturbation parameters to be tolerable, $g : \Theta \times X \to X$ is the image perturbation function, and $F(x) \overset{\text{def}}{=} \underset{1 \le j \le d_O}{\arg\max} f_j(x)$ is the classification function to map an image into the classification label (i.e., the dimension) that corresponds to the maximum output value of DNN $f : X \to \mathbb{R}^{d_O}$. Namely, the property $CR(x;\Theta)$ is satisfied iff the classification label remains the same no matter how image x is perturbed within the range Θ.

Saliency-map. It is known that DNNs classify an input image by paying particular attention to certain specific pixels in the image; a graphical representation of the magnitude of attention to each pixel, like a heatmap, is called *saliency-map* [5]. A saliency-map of each classification label can be obtained from the gradients of each DNN output for an input image [5]; it is defined by the image-to-heatmap function $map_j(x) \overset{\text{def}}{=} \left\langle \frac{\partial f_j(x)}{\partial x_1}, \dots, \frac{\partial f_j(x)}{\partial x_{d_X}} \right\rangle$, where x_i is the i-th pixel value of image x. That is, pixels with higher values of $map_j(x)$ can contribute more to the classification of image x to label j.

Classification Validity. Saliency-maps are used to check the validity of the classification decision basis. For instance, if a DNN classifies the subject type by paying attention to a background rather than the subject to be classified in an input image, it is not a valid basis for classification. We believe such low-validity classification should not be accepted in safety-critical situations, even if the correct classification labels. Semantic perturbations can significantly change the saliency-maps [6], [7]. For example, Fig. 1 shows the changes in MNIST image "8" and the actual saliency-map when the brightness is gradually changed; the collapsed saliency-maps indicate that the DNN does not pay proper attention to text "8" in each image. However, existing robustness verification methods only target changes in the classification labels, not the saliency-maps [8], [9].

II. Our On-Going Work

Attention Robustness. We define the new property of *attention robustness (AR)* — the local robustness of the changes in the saliency-map against combinations of semantic perturbation — by the following predicate:

$$AR(x;\Theta,\delta) \overset{\text{def}}{=} \forall\theta \in \Theta . ai(x;\theta) \le \delta \tag{2}$$

$$ai(x;\theta) \overset{\text{def}}{=} \sum_{j \in Y} \Big\| map_j\big(g(\theta, x)\big) - \tilde{g}\big(\theta, map_j(x)\big) \Big\|_2$$

where $\delta \in \mathbb{R}$ is a given threshold value, $\|\cdot\|_2$ denotes L2-norm, and $\tilde{g} : \Theta \times \mathbb{R}^{d_X} \to \mathbb{R}^{d_X}$ is the heatmap perturbation function that corresponds to the image perturbation function g. Namely, the property $AR(x;\Theta,\delta)$ is satisfied iff $ai(x;\theta)$ remains less than or equal to threshold δ no matter how image x is perturbed within the range Θ. $ai(x;\theta)$ represents *attention inconsistency*, i.e., the difference between **(a)** the actual saliency-map change $map_j\big(g(\theta, x)\big)$ and **(b)** the expected saliency-map change $\tilde{g}\big(\theta, map_j(x)\big)$. Regarding the latter **(b)**, brightness change keeps the saliency-map unchanged, whereas translation moves one along with the image; cf. columns (B) and (T) in Fig. 2. Although the concept of such difference is the same as *saliency-map consistency* used in semi-supervised learning [7], for the sake of verification, it is necessary to calculate the maximum value of ai within Θ.

Verifying AR. Therefore, we propose the first verification method for the property AR. We focus on feed-forward DNNs with rectified linear units (ReLU-FNN) and two facts; **(i)** ReLU-FNN output is linear for its input within each activation region where the activation statuses of all ReLUs are fixed [10] and **(ii)** when x is fixed at $\dot{x}$, many semantic perturbation functions $g(\theta, \dot{x})$ can be interpreted as a ReLU-FNN $g^{\dot{x}}(\theta)$ [4], similarly for $\tilde{g}$. That is, within each activation region $\eta \subseteq \Theta$ of synthesized DNN $f(g^{\dot{x}}(\theta))$, **(a)** the actual saliency-map change $map_j\left(g^{\dot{x}}(\theta)\right)$ is constant because $\frac{\partial f_j(x)}{\partial x_i} = \sum_{s=1}^{d_\Theta} \frac{\partial f_j(g^{\dot{x}}(\theta))/\partial \theta_s}{\partial g_i^{\dot{x}}(\theta)/\partial \theta_s}$ and **(b)** the expected saliency-map change $\tilde{g}^{map_j(\dot{x})}(\theta)$ is a linear function; thus we can efficiently compute $\overline{ai}_\eta^{\dot{x}} \overset{\text{def}}{=} \max_{\theta \in \eta} ai(\dot{x}; \theta)$ by convex optimization and then determine $AR(\dot{x}; \eta, \delta)$ by whether $\overline{ai}_\eta^{\dot{x}} \leq \delta$.

Traversing Outermost Boundary By traversing all activation regions $\eta \subseteq \Theta$ following the existing method [11], we can finally determine $AR(\dot{x}; \Theta, \delta)$ for small ReLU-FNNs. In practice, to represent the trend of the weakness against semantic perturbations for larger ReLU-FNNs, we propose the method to traverse activation regions near the outermost boundary of AR. More formally, the outermost boundary is the connected-space that **(1)** lays on both $AR, \neg AR$ and **(2)** contains the farthest point from the origin $\vec{0} \in \Theta$ through only regions that satisfy AR. The results of traversing the outermost boundaries of CR and AR (cf. Fig. 3); the shapes of the boundaries indicate the existence of regions that satisfy CR but not AR, and the DNN can misclassify the perturbed image with a thin patch and middle brightness in $\neg AR$ regions.

Preliminary Evaluation We implemented our method in a python tool for evaluation. The ReLU-FNN with 200 neurons had a 58% probability of complete verification within 2 hours in the case of traversing all activation regions, and the ReLU-FNN with 800 (2,028) neurons had a 51% (100%) probability of complete verification within 2 hours in the case of traversing outermost boundaries of AR (CR). For further details of our work, please refer to https://arxiv.org/abs/2207.05902.

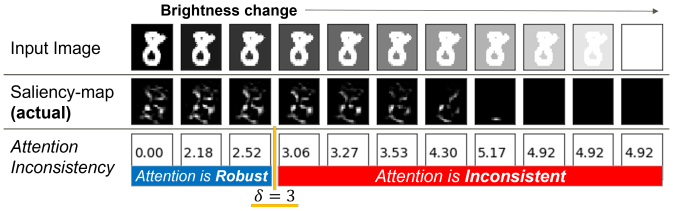

Fig. 1. Perturbation-induced changes in images (first row), the actual saliency-map changes for label 8 (second row), and the values of attention consistency (third row). For threshold $\delta = 3$, the perturbation range up to the third column from the left satisfies attention robustness.

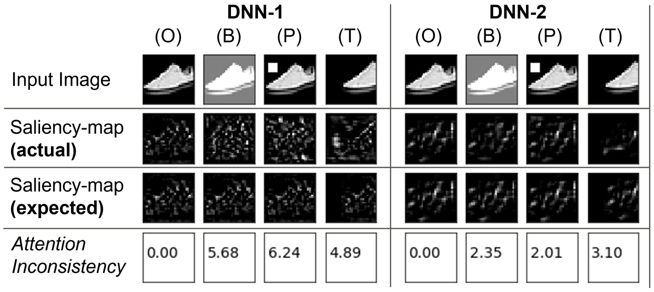

Fig. 2. Differences in changes in saliency-maps for two DNNs. Each saliency-map of DNN-1 above is more collapsed than DNN-2's: where columns (O), (B), (P), and (T) denote original (i.e., without perturbations), brightness change, patch, and translation, respectively.

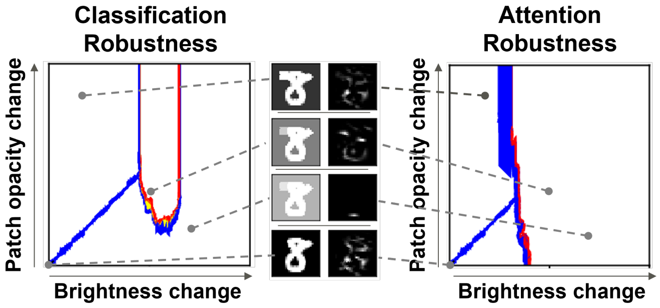

Fig. 3. The results of traversing outermost boundaries of classification robustness (left) and attention robustness (right); blue, red, and yellow regions denote satisfying, fully-violating, and partially-violating each robustness property, respectively. Each plotted point denotes the perturbed input image (middle). The origin at the bottom-left corresponds to the input image without perturbation. Because the starting point of each traversal was the origin, we can see the linear path reaching the outermost boundary.

References

[1] R. Ashmore, R. Calinescu, and C. Paterson, "Assuring the machine learning lifecycle: Desiderata, methods, and challenges," *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–39, 2021.

[2] X. Gao, R. K. Saha, M. R. Prasad, and A. Roychoudhury, "Fuzz testing based data augmentation to improve robustness of deep neural networks," in *ICSE*. IEEE,ACM, 2020, pp. 1147–1158.

[3] M. Balunovic, M. Baader, G. Singh, T. Gehr, and M. Vechev, "Certifying geometric robustness of neural networks," in *NeurIPS*, vol. 32, 2019.

[4] J. Mohapatra, T.-W. Weng, P.-Y. Chen, S. Liu, and L. Daniel, "Towards verifying robustness of neural networks against a family of semantic perturbations," in *CVPR*. IEEE,CVF, 2020, pp. 244–252.

[5] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *ICLR*, 2014.

[6] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[7] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang, "Visual attention consistency under image transforms for multi-label image classification," in *CVPR*. IEEE,CVF, 2019, pp. 729–739.

[8] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Computer Science Review*, vol. 37, p. 100270, 2020.

[9] C. Urban and A. Miné, "A review of formal methods applied to machine learning," *CoRR*, vol. abs/2104.02466, 2021. [Online]. Available: https://arxiv.org/abs/2104.02466

[10] B. Hanin and D. Rolnick, "Deep relu networks have surprisingly few activation patterns," in *NeurIPS*, vol. 32, 2019.

[11] M. Jordan, J. Lewis, and A. G. Dimakis, "Provable certificates for adversarial examples: Fitting a ball in the union of polytopes," in *NeurIPS*, vol. 32, 2019.