Biophysics of adversarial examples

Thomas J. Rademaker[1] and Paul François[1]

***Short Abstract* — Neural networks consistently misclassify adversarial examples, images overlaid with a small and specific perturbation. Similarly, immune cells misclassify agonist ligands in the presence of antagonist ligands in a phenomenon called ligand antagonism. We discovered a mathematical relation between ligand antagonism and adversarial examples, and show how the decision boundary tilts and better approximates the true decision boundary with increasing nonlinearity in both the immune and the neural network.**

***Keywords* — Machine learning, adversarial examples, immune recognition, ligand antagonism, decision boundary**

I. Background

In recent times, neural networks have been immensely successful in performing diverse tasks like object detection, speech recognition, and language translation [1]. Surprisingly, neural nets intrinsically suffer from blind spots, so-called adversarial examples [2]. An imperceptibly small, well-designed perturbation laid over an image will cause the neural net to misclassify the image, while it remains unchanged to the human eye. It has been proposed that adversarial examples are caused by the linearity of neural networks and the high-dimensionality of the data. Indeed, small changes in many pixels can add up to a macroscopic change in the classifier [3]. Others have argued that adversarial examples exist only when the decision boundary lies close to the sampled data, depending on the regularization used during training [4]. At the decision boundary, images are classified with equal probability in either category. At the true decision boundary, images are ambiguous, even for us, whereas at a suboptimal decision boundary we expect to find adversarial examples. It remains an open question on how precisely adversarial effects arise and how more robust neural nets can be designed.

T cells are faced with similar classification tasks as neural networks. They specialize in triggering an immune response when presented with minute amounts of not self ligands while ignoring a vast number of self ligands. Differentiation between ligands is based on the ligand receptor binding kinetics. It is known that antagonist ligands with a dissociation time just below the detection threshold impede the T cell's response to not self ligands via a phenomenon called ligand antagonism [5]. Nature's solution to overcome severe antagonism is to include kinetic proofreading (KPR) and biochemical adaptation in the immune network.

Acknowledgements: This work was funded by a Simons Foundation Investigator Award in the Mathematical Modeling of Living Systems.

[1]Department of Physics, McGill University, 3600 rue University, Montreal, QC H3A 2T8. E-mail: thomas.rademaker@mail.mcgill.ca, paulf@physics.mcgill.ca

We discovered that antagonism in immunology and adversarial examples in machine learning are instances of the same class of problems. Via an analytically tractable model of immune recognition, we established mathematical connections between antagonism and adversarial examples, and explored consequences that until now have been confined to a machine learning context.

II. Results

We applied the Fast Gradient Sign Method [3] to the immune classifier and found that the maximum adversarial perturbation comprises a global decrease of binding times, a decrease in agonist number and an increase in antagonist number, as expected from immunology. Next, we observed when the decision boundary is tilted stronger, the effects of ligand antagonism are weakened, conform [4].

Recent work on Hopfield networks [6,7] discusses the implications of learning with rectified polynomials (RePns), higher order nonlinear activation functions based on Rectified Linear Units. They find their analogue in immune networks via KPR. We showed that prototypic learning is enforced with high order RePns and many KPR steps.

Finally, we demonstrated how networks with higher order RePns or more KPR steps visually and quantitatively better approach the optimal decision boundary. In such networks, adversarial examples and mixtures of antagonists are more robustly classified.

III. Conclusions

Immune networks use proofreading and adaptation to lessen antagonistic effects. Training neural networks with equivalent nonlinear activation functions make them less sensitive to adversarial effects. Our work demonstrates how problems in two very different fields belong to the same class, motivating future studies on the connection between machine learning and biology.

References

[1] LeCun Y, Bengio Y, Hinton GE (2015) Deep learning. *Nature* **521**, 436–444.

[2] Szegedy C, et al. (2013) Intriguing Properties of Neural Networks. arXiv:1312.6199

[3] Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and Harnessing Adversarial Examples. arXiv:1412.6572

[4] Tanay T & Griffin L (2016) A Boundary Tilting Perspective on the Phenomenon of Adversarial Examples. arXiv:1608.07960

[5] François P, Hemery M, Johnson K, Saunders L (2016) Phenotypic Spandrel: Absolute Discrimination and Ligand Antagonism. *Phys. Biol.* **13**, 066011

[6] Krotov D & Hopfield JJ (2016) Dense Associative Memory for Pattern Recognition. *Adv. Neur. Inf. Proc.* **29**, 1172–1180

[7] Krotov, D. and Hopfield JJ (2017) Dense Associative Memory is Robust to Adversarial Inputs. arXiv:1701.00939