Safety Verification for Deep Neural Networks with Provable Guarantees (Extended Abstract)

Marta Kwiatkowska

Department of Computing Science, University of Oxford, UK

Deep neural networks have achieved impressive experimental results in image classification, but can surprisingly be unstable with respect to adversarial perturbations, that is, minimal changes to the input image that cause the network to misclassify it. With potential applications including perception modules and end-to-end controllers for self-driving cars, this raises concerns about their safety. This lecture will describe progress with developing automated verification techniques for deep neural networks to ensure safety of their classification decisions with respect to image manipulations, for example scratches or changes to camera angle or lighting conditions, that should not affect the classification. The techniques exploit Lipschitz continuity of the networks and aim to approximate, for a given set of inputs, the reachable set of network outputs in terms of lower and upper bounds, in anytime manner, with provable guarantees. We develop novel algorithms based on games and global optimisation, and evaluate them on state-of-the-art networks.

Robustness of neural networks is an active topic of investigation and a number of approaches have been proposed to search for adversarial examples. They are based on computing the gradients [1, 3], computing a Jacobian-based saliency map [6], transforming the existence of adversarial examples into an optimisation problem [2], and transforming the existence of adversarial examples into a constraint solving problem [5]. In contrast, this lecture reports on research that aims to rule out the existence of adversarial examples, which approaches based on heuristic search are not able to achieve. In particular, we will adopt the definition of safety based on pointwise robustness introduced in [4], where the first practical automated verification method was developed, based on discretising the neighbourhood and searching it exhaustively in a layer-by-layer manner. A brief overview will also be given of two approaches that utilise Lipschitz continuity, one based on global optimisation [7], and capable of expressing the safety of [4] as well as reachability, and the other [8, 9] on reducing dimensionality by working with black or grey box feature extraction and searching for adversarial examples using a two-player game, where the first player targets the features and the second targets pixels within the feature. The game tree is traversed using Monte Carlo tree search and variants of A* and Alpha-Beta pruning, which produces successive lower and upper bounds on the maximum safe radius with asymptotic convergence guarantees.

References

1. Biggio, B., et al.: Evasion attacks against machine learning at test time. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013. LNCS, vol. 8190, pp. 387–402. Springer, Heidelberg (2013)
2. Nicholas, C., David, W.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE (2017)
3. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. CoRR
4. Huang, X., Kwiatkowska, M., Wang, S., Wu, M.: Safety verification of deep neural networks. In: Majumdar, R., Kunčak, V. (eds.) CAV 2017. LNCS, vol. 10426, pp. 3–29. Springer, Cham (2017)
5. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: an efficient SMT solver for verifying deep neural networks. In: Majumdar, R., Kunčak, V. (eds.) CAV 2017. LNCS, vol. 10426, pp. 97–117. Springer, Cham (2017)
6. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387. IEEE (2016)
7. Ruan, W., Huang, X., Kwiatkowska, M.: Reachability analysis of deep neural networks with provable guarantees. In: International Joint Conference on Artificial Intelligence (2018)
8. Wicker, M., Huang, X., Kwiatkowska, M.: Feature-guided black-box safety testing of deep neural networks. In: Beyer, D., Huisman, M. (eds.) TACAS 2018. LNCS, vol. 10805, pp. 408–426. Springer, Cham (2018)
9. Wu, M., Wicker, M., Ruan, W., Huang, X., Kwiatkowska, M.: A game-based approximate verification of deep neural networks with provable guarantees. CoRR, abs/1807.03571 (2018)