MATHIAS LECUYER, University of British Columbia

Adaptive Randomized Smoothing: Certified Adversarial Robustness for Multi-Step Defences

ML theory usually considers model behaviour in expectation. In practical deployments however, we often expect models to be robust to adversarial perturbations, in which a user applies deliberate changes to on input to influence the prediction a target model. For instance, such attacks have been used to jailbreak aligned foundation models out of their normal behaviour. Given the complex models that we now deploy, how can we enforce such robustness properties while keeping model flexibility and utility?

I will present recent work on Adaptive Randomized Smoothing (ARS), an approach we developed to certify the predictions of test-time adaptive models against adversarial examples. ARS extends the analysis of randomized smoothing using f-Differential Privacy, to certify the adaptive composition of multiple steps during model prediction. We show how to instantiate ARS on deep image classification to certify predictions against adversarial examples of bounded ℓ_∞ norm.