Towards neural networks which know when they don't know

Matthias Hein

Eberhard-Karls-Universität Tübingen, FB Informatik, Tübingen, Germany

Current deep neural networks for image recognition do not know when they don't know. Non-task related images are assigned to one of the classes with high-confidence. Thus the machine learning system cannot trigger human intervention or go over into to a safe state when it is applied out of its specification. I will discuss our recent work to tackle this problem via certifiably adversarially robust out-of-distribution detection.