Ian Goodfellow

ADVERSARIAL MACHINE LEARNING


40

BENCHMARKING TOOLS FOR ADVERSARIAL EXAMPLES

CLEVERHANS

MOST DEFENSES AGAINST ADVERSARIAL EXAMPLES THAT HAVE BEEN PROPOSED SO FAR JUST DO NOT WORK VERY WELL AT ALL, BUT THE ONES THAT DO WORK ARE NOT ADAPTIVE. THIS MEANS IT IS LIKE THEY ARE PLAYING A GAME OF WHACK-A-MOLE: THEY CLOSE SOME VULNERABILITIES, BUT LEAVE OTHERS OPEN.

Ian Goodfellow