An Enhanced Adversarial Attack Method for Defending Against CAPTCHA Solvers

Lai, Bo-Feng (School: Taipei Municipal Jianguo High School)
Wei, Luther (School: Taipei Municipal Jianguo High School)

CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) are often used on websites to prevent users' bots from gaining access to website resources automatically. Yet, lots of researches working on recognizing CAPTCHAs with DNNs (Deep Neural Network) have been proposed and therefore, CAPTCHAs are not considered effective approaches anymore. Thus, to solve this problem, many researchers show that applying adversarial attack techniques to CAPTCHA images is a good way to defend against CAPTCHA solvers. Unfortunately, the general adversarial attack can only be applied on a single well-known model. This is a big challenge for the practical use since each CAPTCHA solver has its own learning model and is unknown to the website owners. A trivial solution is to apply adversarial attacks on all possible models, which increases the interaction times and may become interference for human users. Some researchers try to solve this problem by assembling multiple models together. However, if two models are very different from each other, it would be improper to assemble them together because the adversarial attack for one model may interfere with the other one. In this research, we propose a new adversarial attack method named "Multi-target Ensemble Adversarial Attack". We assume that the CAPTCHA's model candidate list is well-known but the actual recognition model of the attacking solver is unknown. We find the most similar models from the candidate list, assembling them together and applying the adversarial attack on the assembled model. The algorithm is repeated till all models on the candidate list are processed. Experimental results show that our mechanism can decrease the interaction times and block the CAPTCHA solver effectively.