Adversarial Noise for CAPTCHA Solver Protection

2030001 Hiroaki Inoue

The purpose of this study is to analyze the effectiveness, limitations, and problems of CAPTCHA images by providing a means of "defense" against "attacks" that attempt to use machine learning to recognize them. CAPTCHA stands for "Completely Automated Public Turing test to tell Computers and Humans Apart". It is a turing test to distinguish between computers and humans, and is used to prevent spam attacks on Web sites.There are several types of CAPTCHAs, but this study focuses on "Pix," which selects an object with a specified content from among photos.

Advances in image recognition technology have made it easier for computers to break through CAPTCHA tests. Therefore, it is desirable to have a method to make only image recognition models misrecognize images without affecting human vision, perception, and recognition systems. As a defense method, we tested several methods that are likely to inhibit the image recognition model, such as Adversarial Noise, Adversarial Patch, and searching for the weakest class.

And these experiments showed that the method of mixing multiple object classes is effective in inhibiting image recognition models. (178 words)