Research proposal

CAPTCHA : Builds a machine that can read distorted text

Junsu Kim	Seunghyun Kim	Sangho Lee	Eunbyeol Hwang
2014002906	2014002833	2016043509	2018034511
Electronic Engineering	Electronic Engineering	Ecnomics & Finance	Department of Dance

I. Introduction & Motivation

Many online services such as webamil, social media, cloud storage are using CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) as one of major defense mechanisms against abusing bots. However, even though a number of sophisticated CAPTCHA schemes have been suggested, they are suffering from significant vulnerabilities made by common computer vision technologies and macshine learning algorithms. In this paper, we aim to show how CAPTCHA schemes are vulnerable by building a framework that breaks the one of CAPTCHA with well-known techniques.

II. Background

2.1 CAPTCHA A CAPTCHA is an automated test to identify whether the user is a human or not. CAPTCHAs are categorized into two groups. First, text-based CAPTCHA is a commonly used approach to protect web services from bots. The users should identify distorted text and type original letters. Second, image-based CAPTCHA is an another approach that is more user-friendly and easier to solve than conventional text-based CAPTCHAs. In this case, users should choose the image that the service asked

2.2 Machine Learning Algorithms

The decision tree is one of powerful data mining techniques, well-known for its clarity. Decision trees have a tree-like structure in which each internal node decides the direction of decision processes. The result from decision trees can be easily evaluated so that we can quickly identify the key features that affect to the performance of the machine.

The K-Nearest Neighbors algorithm (K-NN) is an algorithm for classification and regression. The learning process includes only of storing all the patterns of the data. During the inference process, the machine calculates the distance between a new input pattern and every saved patterns. As a result, the output becomes the most common value from the k-nearest neighbors.

Support Vector Machines (SVM) is a machine learning algorithm used for classification. It tries to find an optimal hyperplane that divides given feature space. The optimal hyperplane is defined as the hyperplane with a maximum value of minimal distances between the input data and the plane.

Aritificial Neural Network (ANN) has a structure like biological neurons and mimics their interconnections. Each neuron calculates the output from the input data and its own information, then transfers its the output through the network. During the learning phase, the neurons keep modifying its own information called weight, by detecting distance between actual outputs and calculated output within the training phase.