情報保護學會論文誌
第 22 卷 第 4 號, 2012. 8

복수의 이미지를 합성하여 사용하는 캡차의 안전성 검증*

변 제 성,[1†] 강 전 일,[1] 양 대 헌,[1] 이 경 희[2‡]
[1]인하대학교, [2]수원대학교

On the Security of Image-based CAPTCHA using Multi-image Composition*

JeSung Byun,[1†] JeonIl Kang,[1] DaeHun Nyang,[1] KyungHee Lee[2‡]
[1]Inha University, [2]The University of Suwon

요 약

컴퓨터와 사람을 구분하기 위한 수단인 캡차는 광고, 스팸 메일, DDoS 등의 공격을 하는 자동화된 봇을 막기 위해 널리 사용되고 있다. 초창기에는 문자가 출력된 이미지를 왜곡시켜 이를 컴퓨터가 식별하기 어렵도록 하는 방식이 주로 사용되었지만, 이러한 방법들은 인공지능 기법이나 이미지 처리 기법으로 쉽게 무력화 될 수 있음이 여러 연구들을 통해 밝혀졌다. 그러한 이유에서 문자 기반 캡차의 대안으로 이미지를 사용하는 캡차가 주목받게 되었고 그에 따라 여러 가지 형태의 이미지 기반 캡차가 제안되었다. 하지만 텍스트 기반 캡차보다 높은 보안성을 제공하기 위해서는 많은 양의 소스 이미지가 필요하였다. 이에 따라 강전일(2008) 등은 소규모의 이미지 데이터베이스를 이용한 이미지 기반 캡차를 제안하였다. 이 캡차는 사용자 실험을 통해 현재 널리 사용되는 문자 기반 캡차에 비해 사용자 편의성을 보였지만, 아직 안전성이 검증되지 않았다. 이 논문에서는 강전일(2008)등이 제안한 복수의 이미지를 합성하여 사용하는 캡차를 실제로 공격해봄으로써 해당 캡차의 안전성을 검증해 보았다.

ABSTRACT

CAPTCHAs(Completely Automated Public Turing tests to tell Computer and Human Apart) have been widely used for preventing the automated attacks such as spam mails, DDoS attacks, etc.. In the early stages, the text-based CAPTCHAs that were made by distorting random characters were mainly used for frustrating automated-bots. Many researches, however, showed that the text-based CAPTCHAs were breakable via AI or image processing techniques. Due to the reason, the image-based CAPTCHAs, which employ images instead of texts, have been considered and suggested. In many image-based CAPTCHAs, however, the huge number of source images are required to guarantee a fair level of security. In 2008, Kang et al. suggested a new image-based CAPTCHA that uses test images made by composing multiple source images, to reduce the number of source images while it guarantees the security level. In their paper, the authors showed the convenience of their CAPTCHA in use through the use study, but they did not verify its security level. In this paper, we verify the security of the image-based CAPTCHA suggested by Kang et al. by performing several attacks in various scenarios and consider other possible attacks that can happen in the real world.

Keywords: CAPTCHA, Cognitive Security, Support Vector Machine

접수일(2011년 7월 22일), 수정일(1차: 2012년 1월 12일, 2차: 2012년 5월 30일), 게재확정일(2012년 6월 26일)
* 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행되었습니다 (2012-0002146)
† 주저자, tomylove@isrl.kr
‡ 교신저자, khlee@suwon.ac.kr

I. 서 론

캡차(CAPTCHA, Completely Automated Public Turing test to tel Computer and Human Apart)는 컴퓨터와 사람을 식별할 수 있게 해주는 공개된 튜링 테스트를 의미한다[3,7]. 사람은 쉽게 풀 수 있지만 현재의 컴퓨터가 풀기 힘든 인공지능에 관련된 문제들이 캡차에 사용된다. 이러한 특징 때문에 현재 캡차는 홈페이지의 자동 계정 생성, 카페 자동 가입 신청, 블로그 댓글의 스팸광고, 투표 봇, 게시판 글의 자동 등록, DDoS 공격 등을 차단하는데 유용하게 쓰이고 있다[1,22].

그러나 초기 텍스트 기반 캡차의 약점들이 드러나면서 Yahoo의 EZ-gimpy CAPTCHA[2], Widows Live Hotmail CAPTCHA[4], Windows MSN CAPTCHA[5], Google의 Gmail CAPTCHA[6]등이 공격당하게 되었다. 그 대안으로 여러 가지 이미지 기반의 캡차가 만들어졌지만, 사용되지는 않고 있다가 최근에 다시 연구가 이루어지고 있다. 현재까지 알려진 이미지 기반의 캡차로는 카네기 멜론 대학팀의 PIX CAPTCHA[3,7]와 Oli Warner의 Kitten-Auth[9], Chew 와 Tyger의 이미지 기반 캡차[10], Microsoft Research 팀의 Asirra[11]가 있다. 하지만 이미지 기반의 캡차 또한 문제점을 지니고 있다. 이미지 기반 캡차의 보안성은 소스 이미지의 개수에 의존하기 있기 때문에, 충분한 안전성을 제공하기 위해서는 다수의 이미지가 사용되어야 하는데, 이는 이미지 기반 캡차 시스템을 구축하는데 있어 많은 비용이 필요하게 만든다. 따라서 상대적으로 적은 수의 비용에 더 높은 보안성을 갖는 캡차가 필요하게 되었다. 강전일 등은 소규모의 이미지 데이터베이스를 사용하여도 충분한 안전성을 제공할 수 있도록 고안된 복수의 이미지를 합성하여 사용하는 이미지 기반의 캡차를 제안하였다[8]. 이 캡차는 이미지 데이터베이스로부터 무작위로 복수개의 이미지를 선택하여 변형한 후 하나의 이미지로 합치는 방식을 사용하고 있다. 사람의 경우 다수의 이미지를 하나로 합치더라도 하나의 개체 이상을 인식이 가능하지만 컴퓨터의 경우 개체를 인식하기 어렵다는 점을 이용하고 있다. 하지만 해당 논문에서는 하나로 합쳐진 이미지를 사용하는 캡차의 안전성에 대한 검증을 시도하지 않아, 이 논문에서는 복수의 이미지를 합성하여 사용하는 이미지 기반 캡차의 가장 큰 장점인, 적은수의 소스 이미지를 사용하더라도 충분한 안전성을 제공할 수 있다는 점을 검증하기 위해 실제로 공격을 시도해 보았다. 그 결과 적은 수의 소스 이미지를 사용한 경우 충분한 안전성을 제공한다고 할 수는 없지만 기존의 이미지 기반 캡차들에 비해서는 높은 안전성을 제공한다는 사실을 검증할 수 있었다.

이 논문의 구성은 다음과 같다. 2장에서는 기존의 텍스트 기반 및 이미지 기반 캡차들의 종류와 관련된 연구 활동을 살펴보고, 이미지 기반 캡차의 보안 이슈와 안전성을 검증해볼 복수의 이미지를 합성하여 사용하는 이미지 기반의 캡차에 대해 살펴본다. 3장에서는 해당 기법을 실제로 기계 학습을 이용하여 공격을 시도한 결과와 기계학습 이외에 재전송 공격과 인공지능 기법을 활용한 공격, 이미지 처리 기법을 활용한 공격의 가능성에 대한 분석결과를 소개한다. 4장에는 이 논문의 결론을 담는다.

II. 관련연구

2.1 텍스트 기반 캡차

텍스트 기반 캡차는 텍스트를 이용하여 사람과 컴퓨터를 구분하는 기술로써, 무작위로 문자열을 선택한 후, 이를 왜곡된 형태의 이미지로 만들어 사용자에게 해당 이미지에 어떤 문자가 적혀 있는지 물어보는 방식이다. 오랫동안 캡차를 무력화시키기 어렵다고 알려져 왔었지만, 많은 연구들을 통해 캡차가 무력화될 수 있음이 증명되었다.

Mori 와 Malik(2003)는 야후에서 사용되는 Ez-Gimpy 와 Gimpy를 무력화하는데 성공하였다[2]. 그들의 실험에 따르면 Ez-Gimpy는 83~92% 확률로 무력화하는데 성공하였으며, Gimpy는 33% 확률로 무력화할 수 있었다. 또한 Chellapilla 등이 Yahoo나 MSN, Google등에서 일반적으로 사용되는 캡차들이 일반적인 신경망 네트워크나 자신들이 개발한 알고리즘을 이용하면 4.89~66.2% 확률로 무력화할 수 있음을 보였으며[12], 또한 왜곡된 문자는 사람보다는 컴퓨터가 더 잘 인식할 수 있음을 보였다[13,14,15]. 몇몇 프로젝트에서는 캡차를 무력화하기 위한 시도를 하였으며, aiCaptcha 프로젝트는 스팸 봇이 캡차를 무력화하여 블로그에 댓글을 남길 수 있음을 보였으며[16], PWNtcha 프로젝트는 캡차가 무력화된 결과를 보임으로써 시각적인 캡차가 안전하지 않다고 주장하였다[17]. 또한 Marti(2010) 등이 ReCATPCHA가 reCaptchaaOCR을 통해 높은

확률로 무력화될 수 있을 뿐만 아니라, 캡차를 해결해 주는 외부 서비스 업체의 높은 서비스 질에 대해서 조사 분석하였다[23].

이러한 환경 하에서, 기존 캡차의 특징에 새로운 기법을 추가한 캡차가 필요하게 되었다. Chellapilla(2005) 등이 제안한 분할 기반 캡차는 각각의 문자들을 왜곡한 후 추가적으로 원이나 선을 추가한 방식으로 Microsoft가 이러한 방식의 캡차를 사용하고 있다[12].이 또한 Yan 등이 ACM CCS 2008에서 Microsoft 캡차를 60%이상의 확률로 무력화할 수 있다는 것을 보였다[5].

2.2 이미지 기반 캡차

일반적으로 이미지 기반 캡차는 무작위로 선택하고 왜곡한 이미지가 무엇인지 물어보거나, 복수의 이미지에서 동일한 개체의 이름을 물어보는 방식을 사용하고 있다. Ahn(2003) 등이 처음으로 이미지 기반 캡차를 EUROCRYPT 2003에서 제안하였다[3]. Chew 와 Tygar(2004)는 총 3개의 이미지 기반 캡차(the naming CAPTCHA, the distinguish CAPTCHA, the anomalies CAPTCHA)를 제안하였다[10]. Oil Warner의 KittenAuth[9]와 Microsoft Research팀이 제안한 Asirra[11]는 다양한 동물 이미지 중에서 고양이 이미지만 선택하는 방식의 캡차이다. The HotCaptcha[20]는 사용자에게 아름다운 여성의 사진을 선택하게 하는 방식의 캡차이며, HotOrNot.com에서 서비스되고 있다.

이미지 기반 캡차에서는 캡차에 사용될 수 있는 이미지의 개수에 제한이 있다고 하더라도, 가능한 모든 이미지 왜곡의 경우를 컴퓨터에게 학습시키는 것은 불가능하기 때문에 일반적으로 현재 컴퓨터는 이미지를 올바르게 인식할 수 없다고 가정하고 있다.

이미지 기반 캡차의 안전성은 컴퓨터가 모든 이미지에 대해 메타 데이터를 연결할 수 없기 때문에 이미지 데이터베이스 크기에 따라 좌우된다. 그러나 데이터베이스의 크기로 캡차의 안전성을 측정할 수 없다. 예를 들면 봇과 같은 자동화된 에이전트들이 무작위 추측 공격(random guessing attack)을 수행하여 300만개의 이미지를 사용하는 Asirra를 1/4096의 확률로 통과한 전례가 있으며[11], Golle는 SVM(Support Vector Machine)기반 분류기(classifier)를 이용하여 10.3% 확률로 Asirra를 무력화시키는데 성공하였다고 ACM CCS 2008에서 밝혔다[21]. 또한 컴퓨터가 naming CAPTCHA에서 사용자가 해당 테스트 이미지에 어떤 답을 하는지 한차례만 지켜본다면[10], 컴퓨터는 해당 이미지에 대한 메타 데이터를 얻을 수 있고, 약간의 사람의 도움을 얻는다면 naming CAPTCHA를 무력화 시킬 수 있으며, ESP-PIX[3]과 KittenAuth[9]또한 유사한 방식으로 무력화될 수 있다. 기존의 이미지 기반 캡차들의 문제점은 이미지 인식을 어렵게 하는 것이 아닌 이미지를 분류하기 어렵게 하는데 기반을 두고 있다는 것이다. 컴퓨터가 이미지를 인식하는 게 어렵지 않다면, 테스트에 사용되는 몇 개의 이미지를 모은 후, 사람의 도움을 받아 모아진 이미지의 메타정보를 복원할 수 있을 것이다. 컴퓨터는 이렇게 메타정보를 복원한 이미지들이 테스트에 다시 사용될 때까지 여러 번 재시도 끝에 캡차를 무력화 시킬 수 있다. 이에 따라 컴퓨터가 사람의 도움을 받아 메타정보를 복원한 이미지를 가지고 있더라도, 무력화시키기 어려운 새로운 이미지 기반 캡차가 필요하게 되었다.

〔그림 1〕 복수의 이미지를 합성하여 사용하는 이미지 기반의 캡차[8]

2.3 복수의 이미지를 합성하여 사용하는 이미지 기반의 캡차[8]

강전일 등은 복수의 이미지를 합성하여 사용하는 이미지 기반의 캡차를 제안하였다[8]. 이 캡차는 소규모의 데이터베이스를 이용하고 있으며, 이미지 기반 캡차의 안전성을 향상시키기 위해 〔그림 1〕처럼 복수의 이미지를 하나의 이미지로 합성하는 방법을 제안하였다. 이미지 합성에는 여러 가지 방법 중에서 오버레이 방법을 사용하였기 때문에 복수의 이미지를 하나의 이미지로 합성 하였을 때에도 사람은 여전히 하나 이상의 개체에 대하여 인식을 수행할 수 있다. 또한 이미지를 어떠한 변형 없이 하나의 이미지로 합성한 경우, 복수의 이미지가 각자의 이미지로 분리될 수 있기 때문에, 합성하기 전에 이미지를 각각 변형한 후 하나의 이미지로 합치는 방식을 사용하였다. 강전일 등이 제안한 기법은 사용자 실험을 통해 기존의 텍스트 기반 캡차에 비해 평균 입력 시간이 짧으며, 평균 입력 타수 또한 기존의 텍스트 기반 캡차보다 적어 기존의 텍스트 기반 캡차에 비해 사용자 편의성 측면에서의 우수함을 보였다. 하지만 해당 논문에서는 합성된 이미지 자체의 안전성이 검증하지 않았다. 이 논문에서는 복수의 이미지를 합성하여 사용하는 캡차를 기계 학습을 통해 실제로 공격해 봄으로써 해당 논문에서 제안한 기법의 안전성을 측정해 보았다.

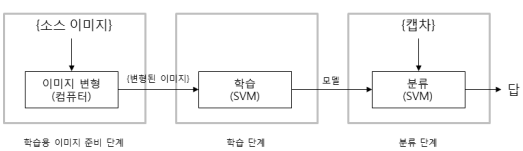
III. 기계 학습(Machine learning)을 통한 공격 및 보안성 분석

3.1 일러두기

이 논문에서는 〔표 1〕과 같은 기호를 사용한다. 또한 〔표 2〕와 같은 실험 환경에서 실험을 진행 하였다.

3.2 기계 학습(Machine learning)을 통한 공격

이 논문에서는 강전일 등이 제안한 기법의 안전성



〔그림 2〕 시나리오 1의 실험 방법

〔표 1〕 논문에서 사용되는 기호 및 설명

기호	설명
$\mathbb{S}_i$	소스 이미지 집합 $\mathbb{S}$의 i번째 소스 이미지
$\mathbb{C}$	복수의 이미지를 합성한 캡차 이미지
$\mathbb{T}_i = \{\mathbb{I}_{i,k}\}$	레이블이 i인 SVM 학습용 이미지 집합
$\mathbb{I}_i$	임의로 변형된 중간 이미지
$\mathbb{I}_i \leftarrow F^n(\mathbb{S}_i)$	소스 이미지 $\mathbb{S}_i$를 임의의 n가지 필터를 적용하여 새로운 이미지 $\mathbb{I}_i$를 생성하는 함수 ($n=1$인 경우는 1생략)
$\mathbb{C} \leftarrow M(\mathbb{I}_i, \mathbb{I}_j, \cdots)$	복수의 이미지 $\mathbb{I}_i$, $\mathbb{I}_j$ 등을 합성 후, 노이즈를 추가하여 캡차를 생성하는 함수
$(x, y, \cdots) \leftarrow R(\mathbb{C})$	캡차로부터 레이블(해답)을 추출하는 가상 함수 (사람이 수행)

〔표 2〕 실험 환경

실험환경	
CPU	Intel Xeon Nocona@3.0 Ghz
메모리	8Gb
운영체제	Ubuntu Server 11.04
커널	3.0.0-19 server
구현 언어 및 라이브러리	Perl 5.10, Imager, GD, LIBSVM[24], LIBLINEAR[25]

을 확인하기 위해, 복수의 이미지를 합성하여 사용하는 이미지 기반의 캡차에 대해 두 가지의 서로 다른 가정 아래에서 공격을 시도하였다. 두 가지의 경우에 대해서 실험을 진행하였는데 첫 번째 경우는 컴퓨터(공격자)가 강전일 등이 제안한 캡차에 사용되는 200×200 픽셀의 소스 이미지 200개를 모두 가지고 있다는 가정하고 실험을 진행하였으며, 또 다른 경우는 공격자가 소스 이미지를 가지고 있지 않다는 가정하에 실험을 하였다. 실험을 위해서 이 논문 에서는 LIBSVM[24]과 LIBLINEAR[25]를 사용하였으며, 실험 환경은 〔표 1〕과 같다.

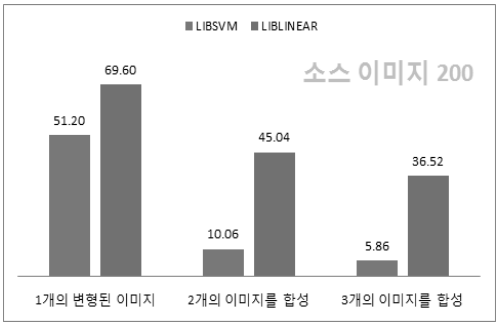
3.2.1 시나리오 1 : 공격자가 모든 소스 이미지를 가지고 있는 경우

강전일 등이 제안한 기법에서는 각각의 소스 이미지를 변형한 후에 하나로 합치는 방식을 사용하고 있다. 모든 소스 이미지를 공격자가 가지고 있다고 하더라도, 각각의 소스 이미지가 그대로 캡차 테스트에 사용되지 않는다.

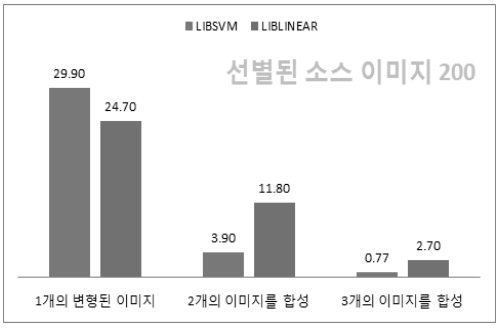
〔그림 2〕는 시나리오 1의 실험 방법을 간략히 보여주고 있다. 우선, 공격자는 자신이 보유하고 있는 소스 이미지 집합 $\mathbb{S}$으로부터 학습용으로 사용할 이미지를 준비한다. 어떠한 소스 이미지 $\mathbb{S}_i$이미지로부터 $\mathbb{I}_{i,k} \leftarrow F^2(\mathbb{S}_i)$처럼 학습용 20개로 이루어진 학습 이미지 집합 $\mathbb{T}_i = \{\mathbb{I}_{i,k}\}_{k=1..20}$를 만든다. 공격자는 200개의 소스 이미지로부터 4,200개의(=(20+1)×200)) 테스트 이미지를 생성하고 SVM을 통해 학습 시킨다. 공격자는 학습된 결과를 이용하여, 1개의 소스 이미지를 변형시킨 경우(즉, $\mathbb{C}=F^2(\mathbb{S}_i)$)와 2개의 소스 이미지를 합성하여 사용하는 경우(즉, $\mathbb{C}=M(F^2(\mathbb{S}_i),F^2(\mathbb{S}_j))$), 마지막으로 3개의 소스 이미지를 합성하여 사용하는 경우(즉, $\mathbb{C}=M(F^2(\mathbb{S}_i),F^2(\mathbb{S}_j),F^3(\mathbb{S}_k))$에 대해서 분류를 시도하여 보았다.

또한 소스 이미지의 선택이 캡차에 미치는 영향에 대해 알아보기 위하여, 배경이 단색으로 이루어진 이미지를 선별하여, 배경이 다수의 색상으로 이루어진 이미지로 대체한 후, 200개의 소스 이미지로부터 4200개의(=(20+1)×200)) 테스트 이미지와, 500개의 소스 이미지로부터 10500(=(20+1)×500)) 테스트 이미지를 생성하였다. 이렇게 생성된 테스트 이미지들을 SVM을 통해 학습 시킨 후, 1개의 소스 이미지를 변형시킨 경우와 2개의 소스 이미지를 합성하여 사용하는 경우, 마지막으로 3개의 소스 이미지를 합성하여 사용하는 경우에 대해서 분류를 시도하여 보았다.

(실험 방법)



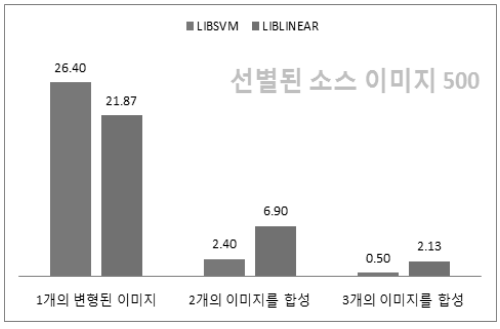
〔그림 3〕 200개의 소스 이미지를 선별하지 않은 경우의 시나리오 1의 실험 결과



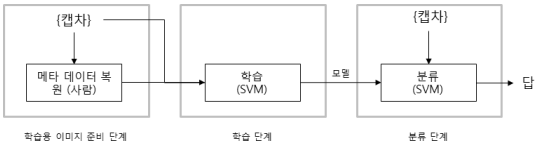
〔그림 4〕 200개의 소스 이미지를 선별한 경우의 시나리오 1의 실험 결과 (단위: %)

(실험 결과)

선별되지 않은 200개의 소스 이미지를 사용한 경우의 실험 결과는 〔그림 3〕에서 보여지는바와 같이 1개의 소스 이미지만 변형시킨 경우(기존의 이미지 기반 캡차)에는 약 70%의 확률로 분류하는데 성공하였으며, 2~3개의 이미지를 합성 하여 사용하는 경우[8]에는 약 36~45%의 확률로 분류하는데 성공하였다. 하지만 선별된 200개의 소스 이미지를 사용한 경우의 실험 결과는 〔그림 4〕에서 보여지는바와 같이 1개의 소스 이미지만 변형시킨 경우(기존의 이미지 기반 캡차)에는 약 30%의 확률로 분류하는데 성공하였으며, 2~3개의 이미지를 합성 하여 사용하는 경우에는 약 2~12%의 확률로 분류하는데 성공하였으며, 선별된 500개의 소스 이미지를 사용한 경우의 실험 결과는 〔그림 5〕에서 보여지는바와 같이 1개의 소스 이미지만 변형시킨 경우(기존의 이미지 기반 캡차)에는 약 26%의 확률로 분류하는데 성공하였으며, 2~3개의 이미지를 합성 하여 사용하는 경우에는 약 2~7%의 확률로 분류하는데 성공하였다. 이는 소스 이미지의 배경이 단색으로 이루어진 경우, 정보의 엔트피가 낮아져, 캡차의 안정성에 영향을 미친다는 사실을 보여주고 있다.



〔그림 5〕 500개의 소스 이미지를 선별한 경우의 시나리오 1의 실험 결과 (단위: %)



〔그림 6〕 시나리오 2의 실험 방법

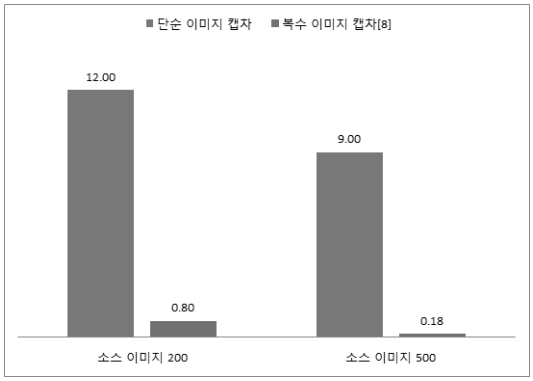
3.2.2 시나리오 2 : 공격자가 소스 이미지를 가지고 있지 않는 경우

소스 이미지를 가지고 있지 않는 공격자는 복수의 이미지를 합성하여 사용하는 이미지 기반의 캡차를 공격하기 위해서는 캡차 시스템에 자주 접근하여 많은 수의 테스트 이미지를 모은 후 각각의 이미지들에 메타 데이터를 입력하여야한다. 컴퓨터는 이미지를 인식하여 스스로 메타 데이터를 복원할 수 없기 때문에 협력자(사람)의 도움이 반드시 필요하다.

(실험 방법)

〔그림 4〕는 시나리오 2의 실험 방법을 간략히 보여주고 있으며, 자세한 실험 방법은 다음과 같다. 먼저 실험을 위해 무작위로 선정된 2개의 소스 이미지 $\mathbb{S}_i, \mathbb{S}_j$를 변형 함수 $F^2()$를 두 번 적용해 변형 시킨 후 하나로 이미지로 합친 캡차(즉, $\mathbb{C}=M(F^2(\mathbb{S}_i),F^2(\mathbb{S}_j))$)를 학습용 이미지로 준비하였다. 그 중 ℓ개의 학습에 적합하도록 선택된 캡차에 대하여[1] 다음과 같이 학습용 이미지 집합을 준비한다. 수집된 캡차 $\mathbb{C}_i$는 사람의 도움을 받아 $(x,y)\leftarrow R(\mathbb{C}_i)$처럼 레이블을 복원한 뒤, 복원된 레이블을 이용하여 학습용 이미지 집합 $\mathbb{T}_x$와 $\mathbb{T}_y$에 $\mathbb{C}_i$를 각각 포함한다. 모든 캡차 이미지에 대하여 레이블을 추출하고 학습용 이미지 집합들 $\mathbb{T}_1,\mathbb{T}_2,\cdots,\mathbb{T}_m$을 이용하여 학습시켰다. 여기서 n은 소스 이미지의 수이고, ℓ은 인간 협력자가 $R()$를 수행한 횟수이다. 이 실험에서 $n=200$에 대해서 $\ell=\{100,200,400,800\}$이고, $n=500$에 대해서 $\ell=\{250,500,1000,2000\}$이다. 서로 다른 n에 대해서 ℓ이 다른 이유는, 학습용 이미지 집합에 하나의 소스 이미지가 포함된 캡차가 출현하는 비율을 일정하게 맞춰주기 위해서이다.

1) 실제 환경에서 공격자(컴퓨터)는 협력자(사람)이 레이블을 구하기 전에는 캡차들은 취사선택할 수 없다. 실험의 목적은 어디까지나 안전성 분석에 초점을 맞추었으므로 안정적인 실험 결과를 위하여, 소스 이미지가 노출이 균등하게 되도록 조정하였다. 실제 협력자의 노력은 쿠폰 수집가 문제에 따라 ℓH_ℓ에 달하며, 이 경우 공격 성공률은 약간 더 높을 수 있다.



〔그림 7〕 시나리오 2에서의 단순 이미지 캡차와 복수 이미지 캡차의 비교 결과(LIBSVM, 단위: %)

(실험 결과)

실험 결과 〔그림 7〕에서 보이는 바와 같이 1개의 소스 이미지만 변형시킨 경우(기존의 이미지 기반 캡차) 9~12%의 확률로 캡차를 분류하는데 성공하였고, 2개의 소스 이미지를 합성한 경우[8]에는 공격자의 동일한 노력 아래에서 0.1~0.8%의 확률로 캡차를 분류하는데 성공하였다. 〔그림 9〕는 복수의 이미지를 합성하여 사용하는 이미지 기반의 캡차[8]에서 협력자에 노력과 소스 이미지의 개수에 따른 공격 성공 확률을 비교한 것이다. 소스 이미지의 개수가 200개에서 500개로 증가할 때, 성공 확률은 낮아졌으며, 소스 이미지를 선택하여 사용하는 것에는 크게 영향을 미치지 않았다. 이 결과로부터, 협력자의 노력만큼의 이득을 취하기 위해서는 200개의 소스 이미지를 사용하는 캡차의 경우 캡차를 대략 4만번 정도 풀려고 시도해야하며(학습용 이미지 1,000개 기준 2.5% 성공률 가정, 1,000개 캡차를 풀면 25개 성공, 따라서 캡차 1,000개를 성공적으로 풀기 위해서는 $40000=1000/0.025$개 시도 필요), 500개의 소스 이미지를 사용하는 캡차의 경우, 캡차를 28만 5천번($=1000/0.0035$)정도 풀려고 시도해야 한다는 사실을 쉽게 유추할 수 있다. 200개의 소스 이미지의 경우 어느 정도 현실성이 있지만, 500개의 소스 이미지의 경우 현실성이 없다고 볼 수 있을 것이다.

시나리오 1과 시나리오 2를 통해서 복수의 이미지를 합성하여 사용하는 캡차가 하나의 이미지를 변형하여 사용하는 캡차보다 보안성 측면에서 우수하다는 것을 알 수 있었다. 그러나 위와 같은 실험환경에서는 복수의 이미지를 합성하여 사용하는 캡차 또한 소수의 소스 이미지만을 사용할 경우 충분한 보안성을 가지고 있다고 말하기는 힘들다.

3.3 다른 다양한 공격의 가능성

3.3.1 재전송 공격

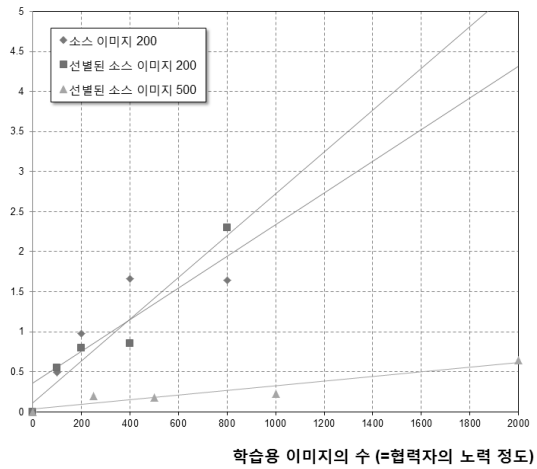
공격자는 수차례 정상적인 사용자가 캡차 시스템에 접근하는 것을 관찰하여 테스트 이미지와 해당 테스트 이미지에 대응하는 메타 데이터를 모은 후, 캡차 시스템에 접근하여 축적한 테스트 이미지가 보일 때 까지 기다린 후 메타 데이터를 입력하여 캡차 테스트를 통과하는 재전송 공격이 가능하다. n개의 소스 이미지에서 m개를 골라 하나의 이미지로 합치는 방식의 경우에는 축적한 테스트 이미지가 캡차에서 다시 나올 확률은 $1/n^m$이지만, 각각의 이미지를 변형 후 하나의 이미지로 합치는 방식의 경우에는 이미지를 왜곡하는 방식에 따라 수천가지의 다른 변형 방법이 존재하기 때문에 공격자가 재전송 공격을 수행하기 해서는 더 많은 노력이 필요할 것이다.

3.3.2 인공지능 기법을 이용한 공격

복수의 이미지를 합성하여 사용하는 이미지 기반의 캡차 시스템[8]에서 사용되는 모든 소스 이미지가 공격자에게 드러난다면, 공격자는 이 소스 이미지를 이용하여 캡차 시스템을 무력화시키기 위한 시도를 할 수 있다. 하지만 공격자는 캡차 시스템에서 사용되는 모든 소스 이미지를 모으기는 어려우며, 모든 소스 이미지를 모으더라도 해당 소스 이미지에 대한 메타 데이터 복원해야 하는 작업 또한 필요하다. 모든 소스 이미지 n개를 수집하는 것은 쿠폰 수집 문제(coupon collector's problem)에 해당한다. 공격자는 모든 소스 이미지 n개를 수집하기 위하여 캡차 시스템에 nH_n번 접근을 시도해야 한다. 만약 2,000개의 소스 이미지를 모으길 원한다면 공격자는 대략 23,000번 캡차 시스템을 접근하여야 한다. 또한 모여진 소스 이미지들의 메타 데이터를 복원해야하는데, 사람에 따라 같은 소스 이미지를 보고 서로 다른 메타 데이터를 복원해내기 때문에, 단순히 캡차 시스템 접근을 통해 축적한 이미지들과 사람을 통해 복원된 메타 데이터는 인공지능 기법의 학습용으로는 적합하지 않을 수 있다.

3.3.3 이미지 처리 기법을 이용한 공격

Szeliski 등이 다수의 이미지가 하나로 이미지로 합성된 경우에 대해서 각각의 이미지들을 분리해내는 기법을 이용하여 복수의 이미지를 합성하여 사용하는 이미지 기반의 캡차[8]를 공격을 할 수 있다. Szeliski가 제안한 기법을 통해 공격하기 위해서는 서로 같은 소스 이미지를 가지고 있지만 서로 다른 변형 함수가 적용된 테스트 이미지가 다수 필요하다. 공격자는 한 개의 소스 이미지를 추출해내기 위해 평균 ${}_nC_m$개의 테스트 이미지가 필요하며, 모든 소스 이미지를 추출하기 위해서는 $(2\cdot {}_nC_m)\times H_{(2\cdot {}_nC_m)}$개의 테스트 이미지가 필요하고, 추가적으로 모은 테스트 이미지를 인공지능 기법의 학습용으로 사용하기 위해서는 사람의 도움을 받아 ${}_nC_m$개의 클래스로 분류해야 한다. 예를 들어 총 2000개의 소스 이미지를 가지고 있으며 무작위로 2개의 소스 이미지를 선택하고 변형 후 하나의 이미지로 합치는 캡차 시스템을 공격하기 위해서 공격자(컴퓨터)는 약 3.59×10^{14}개의 테스트 이미지를 모아야 하며, 모아진 테스트 이미지를 협력자(사람)의 도움을



〔그림 8〕 시나리오 2에서의 협력자의 노력에 따른 성공 확률 비교 (테스트 이미지는 각각 1000개, LIBSVM[24] 사용) (단위: %)

〔표 3〕 다른 다양한 공격 가능성 비교

	재전송	인공지능	이미지 처리
특성	공격하기 위해 적은 노력 필요	공격하기 위해 많은 노력 필요	공격하기 위해 많은 노력 필요
	$1/n^m$ 확률로 성공	기계 학습 기법과 비슷한 성공률	기계 학습 기법과 비슷한 성공률
	협력자의 도움 필요 없음	협력자의 도움 필요	협력자의 도움 필요

받아 1,999,000개의 클래스로 분류 하여야 한다. 캡차 시스템에서 사용되는 소스 이미지의 개수와 합성되는 이미지 개수를 늘린다면 공격자는 더 많은 양의 테스트 이미지를 모아야 하고, 더 많은 클래스로 분류해야 하는 등의 노력이 필요하기 때문에 Szeliski 등이 제안한 방법을 사용하더라도 기계 학습을 통한 공격보다는 이점을 얻을 수 없을 것으로 예상된다.

IV. 결론

이 논문에서는 강전일 등이 제안한 복수의 이미지를 합성하여 사용하는 이미지 기반 캡차[8]의 안전성을 측정하기 위해 200개의 소스 이미지로 구성된 시스템을 실제로 기계 학습을 통해 공격을 시도해보았다. 그 결과 200개의 소스 이미지를 가지고 있는 경우에는 약 36~45%의 확률로 공격에 성공하였으며, 소스 이미지를 가지고 있지 않는 경우에는 공격자의 노력에 따라 차이가 존재하지만 0.5~1.7%의 확률로 공격에 성공할 수 있음을 보였다. 또한 소스 이미지의 선택이 복수의 이미지를 합성하여 사용하는 이미지 기반 캡차에 어떠한 영향을 미치는지 알아보기 위하여, 선별된 200개의 소스 이미지와 500개의 선별된 소스 이미지로 구성된 시스템을 기계 학습을 통해 공격을 시도해보았다. 그 결과 200개의 식별된 소스 이미지를 가지고 있는 경우에는 약 2~12%의 확률로 공격에 성공하였으며, 소스 이미지를 가지고 있지 않는 경우에는 0.6~2.3%의 확률로 공격에 성공할 수 있음을 보였으며, 500개의 식별된 소스 이미지를 가지고 있는 경우에는 약 2~7%의 확률로 공격에 성공하였으며, 소스 이미지를 가지고 있지 않는 경우에는 0.2~0.7%의 확률로 공격에 성공할 수 있음을 보였다. 하지만 현실적으로 공격자는 캡차 시스템의 이미지 데이터베이스를 공격하지 않고서는 모든 소스 이미지를 얻는 것은 불가능하기 때문에, 소스 이미지를 모두 가지고 있는 경우의 공격 결과 보다는 소스 이미지를 가지고 있지 않고 캡차 시스템에 접속하여 소스 이미지를 모으는 후자의 경우가 해당 캡차의 보안 수준을 보여준다고 할 수 있을 것이다.

공격자는 소스 이미지를 모으기 위해서는 여러 번 캡차 시스템에 접속하여야 한다. 이러한 공격자의 행동은 캡차의 관리자에게 쉽게 관측될 수 있을 것이며, 테스트 이미지를 모으더라도 해당 테스트 이미지에 대한 메타 데이터를 복원해야 하는 노력도 필요할 것이며, 복원 메타 데이터 또한 올바른 메타 데이터가 아닐 경우에는 성공 확률 또한 낮아질 것이다.

따라서 해당 캡차 기법이 충분한 안전성을 확보하기 위해서는 선별된 소스 이미지가 수 천개이상의 소스 이미지가 사용 되어야 할 것으로 보인다.

참고문헌

[1] L. Ahn, B. Maurer, C. McMillen, D. Abraham and M. Blum, "reCAPTCHA: Human-Based Character Recognition via Web Security Measures", SCIENCE, vol. 321, no. 5895, pp. 1465-1468, Sep. 2008.

[2] G. Mori and J. Malik, "Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA," Proceedings of the Computer Vision and Pattern Recognition (CVPR) Conference, pp. 134-141, Jun. 2003.

[3] L. Ahn, M. Blum, N. J.Hopper, and J. Langford, "CAPTCHA: Using Hard AI Problems For Security", EUROCRYPT'03, LNCS 2656, pp. 294-311, 2003.

[4] S. Prasad, "Microsoft Live Hotmail Under Attack by Streamlined Anti-CAPTCHA and Mass-mailing Operations," http://securitylabs.websense.com/content/Blogs/3063.asp, 2008.

[5] J. Yan and A.S. El Ahmad, "A Low-cost Attack on a Microsoft CAPTCHA," Proceedings of the 15th ACM Conference of Computer and Communications Security, pp. 543-554, Oct. 2008.

[6] S. Prasad, "Google's CAPTCHA busted in recent spammer tactics," http://securitylabs.websense.com/content/Blogs/2919.aspx, 2008.

[7] L. Ahn, M. Blum, and J. Langford, "Telling Humans and computers apart automatically", Communications of the ACM, vol. 47, no. 2, pp. 56-60, Feb. 2004.

[8] 강전일, 맹영재, 김군순, 양대헌, 이경희, "복수의 이미지를 합성하여 사용하는 이미지 기반의 캡차와 이를 위한 안전한 운용 방법," 정보보호학회논문지, 18(4), pp. 153-165, 2008년 8월.

〔9〕 O. Warner, "KittenAuth Project" http://www.thepcspy. com/ kittenauth.
〔10〕 M. Chew and J. D. Tygar, "Image Recognition CAPTCHAs", Proceedings of the 7th International Information Security Conference (ISC 2004), LNCS 3225, pp. 268-279, 2004.
〔11〕 J. Elson, J. R. Douceur, J. Howell, and J. Saul, "Asirra: a CAPTCHA that exploits interest-aligned manual image categorization", Proceedings of 14th ACM Conference on Computer and Communications Security, pp. 366-374, Nov. 2007.
〔12〕 K. Chellapilla and P. Y. Simard, "Using machine learning to break visual human interaction proofs (hips)," Advances in Neural Information Processing Systems 17 (NIPS'2004), MIT Press, 2004.
〔13〕 K. Chellapilla, K. Larson, P. Y. Simard, and M. Czerwinski, "Computers beat humans at single character recognition in reading based human interaction proofs (hips)," Proceedings of International Conference on Email and Anti-Spam (CEAS), Jul. 2005.
〔14〕 K. Chellapilla, K. Larson, P. Y. Simard, and M. Czerwinski, "Designing human friendly human interaction proofs (hips)," Proceedings of the 2004 ACM Conference on Human Factors in Computing Systems (CHI), pp. 711 - 720, Apr. 2005.
〔15〕 K. Chellapilla, K. Larson, P. Y. Simard, and M. Czerwinski, "Building segmentation based human-friendly human interaction proofs (hips)," Proceedings of Human Interactive Proofs (HIP), LNCS 3517, pp. 1 - 26, 2005.
〔16〕 aiCaptcha, "Using ai to beat captcha and post comment spam,", http://www.brains-n-brawn.com/aiCaptcha
〔17〕 S. Hocevar, "Pwntcha - captcha decoder," http://sam.zoy.org/ pwntcha/
〔18〕 M. May, "Inaccessibility of captcha - alternatives to visual turing tests on the web," W3C Working Group Note 23, http://www.w3.org/TR/turingtest/, Nov. 2005.
〔19〕 M. Uschold and M. Grüninger, "Ontologies: Principles, Methods and Applications," Knowledge Engineering Review, vol. 11, no. 2, pp. 93 - 136, Jun. 1996.
〔20〕 HotCaptcha, http://www.hotcaptcha.com/
〔21〕 P. Golle, "Machine Learning Attacks Against the Asirra CAPTCHA," Proceedings of 15th ACM Conference on Computerand Communications Security(CCS), pp. 535 - 542, 2008.
〔22〕 A. Juels and J. Brainard. "Client Puzzles: A Cryptographic Countermeasure Against Connection Depletion Attacks," Proceedings of the 1999 Networks and distributed system symposium, pp. 151-165 , Feb. 1999.
〔23〕 M. Motoyama, K. Levchenko, C. Kanich, and D. McCoy. "Re: CAPTCHAs: understanding CAPTCHA-solving services in an economic context", Proceedings of the 19th USENIX conference on Security, pp. 435 - 452, Aug. 2010
〔24〕 C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," http://www.csie.ntu.edu.tw/ ~cjlin/libsvm, 2001.
〔25〕 R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. "LIBLINEAR: A library for large linear classification," Journal of Machine Learning Research, vol. 9, pp. 1871-1874, Jun. 2008.

〈著者紹介〉

변 제 성 (JeSung Byun) 학생회원
2009년 2월: 공주대학교 컴퓨터 공학과 졸업
2010년 3월~현재: 인하대학교 컴퓨터 정보 공학부 석사 과정
〈관심분야〉 무선 센서 네트워크 보안, 무선 인터넷 보안, 인증 프로토콜, 웹 인증 보안

강 전 일 (Jeonil Kang) 학생회원
2003년 2월: 인하대학교 컴퓨터 공학과 졸업
2006년 2월: 인하대학교 정보통신대학원 석사
2006년 3월~현재: 인하대학교 정보공학과 박사 과정
〈관심분야〉 RFID 보안, 생체 인식 보안, 무선 센서 네트워크 보안, 무선 인터넷 보안, 웹 인증 보안

양 대 헌 (DaeHun Nyang) 정회원
1994년 2월: 한국과학기술원 과학기술 대학 전기 및 전자 공학과 졸업
1996년 2월: 연세대학교 컴퓨터 과학과 석사
2000년 8월: 연세대학교 컴퓨터 과학과 박사
2000년 9월~2003년 2월: 한국전자통신연구원 정보보호연구본부 선임연구원
2003년 2월~현재: 인하대학교 컴퓨터 정보 공학부 부교수
〈관심분야〉 암호 이론, 암호 프로토콜, 인증 프로토콜, 무선 인터넷 보안

이 경 희 (Kyunghee Lee) 정회원
1993년 2월: 연세대학교 컴퓨터과학과 학사
1998년 8월: 연세대학교 컴퓨터과학과 석사
2004년 2월: 연세대학교 컴퓨터과학과 박사
1993년 1월~1996년 5월: LG소프트(주) 연구원
2000년 12월~2005년 2월: 한국전자통신연구원 선임연구원
2005년 3월~현재: 수원대학교 전기공학과 조교수
〈관심분야〉 바이오인식, 정보보호, 컴퓨터비전, 패턴인식