Innovation Academy X Superman Company*

* 상기 업체는 가상의 업체임

CAPTCHA-2

CAPTCHA는 사람은 인식하지만, 컴퓨터는 인식할 수 없는 이미지를 제시하고

사용자가 사람일 때만 맞추도록 함으로써 컴퓨터에 의한 자동 로그인을 막는 장치로 많이 사용된다.

그러나, 인공지능/머신러닝 시대가 오지 않았는가?

머신러닝을 이용하여 CAPTCHA 이미지를 인식하여 정답을 맞추는 프로그램을 만들어보자.

Version 007 : Release Note
Authors
• Lee, Minsuk, mail@innoaca.com, Innovation Academy
• AI Superpower Lab, Superman Company

Contents

Innovation Academy X Superman Company

프로젝트 스토리

CAPTCHA의 한계

- reCAPTCHA(Original)

- CAPTCHA와 인공지능

이 프로젝트가 해결하고자 하는 문제

프로젝트의 기술적 설명

결과물에 관한 설명

프로젝트를 구현할 때 제약조건

프로젝트 결과의 검증

프로젝트 데이터 파일 설명

학습 지원 정보

프로젝트 수행/리뷰/평가/멘토링 Forum

※ 본 예시에서 사용하는 **'프로젝트'**는 제안요청서(3page) 내
<콘텐츠 구조 및 구성요소 예시> 의 문제 SET 구조 중 **'미션'**에 해당함

프로젝트 스토리

CAPTCHA는 Completely Automated Public Turing test to tell Computers and Humans Apart의 약자로 사용자가 사람인지 컴퓨터인지를 구분하기 위한 방법이다.

(CAPTure(d) + CHAracter, 에서 따왔다고도 한다.) CAPTCHA는 사람은 인식하지만, 컴퓨터는 인식할 수 없는 이미지를 제시하고, 사용자가 사람인 경우에만 답을 맞추도록 함으로써 컴퓨터(봇)에 의한 자동 로그인을 막는 방법으로 많이 사용된다. CAPTCHA는 광고성 게시물 방지, 아이디 자동생성 방지, 이메일 주소 보호, 온라인 선거, 계정 해킹 방지 등에 사용되며, 자동 프로그램으로 부당한 이득을 취하거나 악용하는 사람들, 불필요한 게시물 도배 등을 막을 수 있다.

CAPTCHA의 한계

이미 문자 기반의 CAPTCHA 중 일부는 연구자들에 의해 뚫리고 있으며, 수학 문제 CAPTCHA, 오디오 CAPTCHA 등 다른 CAPTCHA 방식들도 머신러닝 기술의 발전으로 점점 뚫리고 있다. 이를 보완하기 위하여 reCAPTCHA, noCAPTCHA와 같은 다른 방법들 역시 개발되고 있다. 문자 이미지가 아닌 거리, 사물 이미지 기반의 CAPTCHA는 특정 문화에 익숙해져 있지 않으면 맞히기 어려운 문제를 낼 수도 있어 잘 정의된 타켓 사용자만 통과할 수 있도록 설계된다. 또 이미지나 수학 문제 풀기 등에 의존하지 않고, 마우스의 이동 패턴 등을 추적하여 진짜 사람인지를 판별하는 변형된 방식도 사용한다.

reCAPTCHA(original)

(한글 wikipedia에서 전제) 오래 전에 제작된 종이 책들을 텍스트화하기 위해 OCR 프로그램을 사용하는데, 낙서나 얼룩, 헤짐 등의 방해요소만 있어도 OCR 프로그램은 텍스트를 제대로 인식하지 못한다. 이런 단어들은 사람이 하나하나 판독해야 하지만 수요가 적은 책들까지 일일이 입력하려면 노동력과 인건비가 많이 들어간다. 이를 해결하기 위해 CAPTCHA를 입력하는 수 많은 사용자들의 힘을 빌리는 것이 바로 최초 버전의 reCAPTCHA이다. reCAPTCHA는 다음과 같은 과정을 거친다.

reCAPTCHA는 두 개의 암호코드를 제시한다. 하나는 컴퓨터가 이미 답을 알고 있는 단어이지만 다른 하나는 서적을 텍스트화 하는 도중 OCR이 인식하지 못한 단어이다.

사용자가 두 단어를 모두 입력하면 컴퓨터는 이미 답을 아는 단어로 대상이 사람인지를 확인한다.

대상이 사람이라고 판정된 경우 나머지 단어 또한 정답으로 판단한다.

이렇게 인식하지 못한 한 단어를 여러 사람에게 반복해서 테스트한 뒤 높은 비율로 입력된 단어를 선택해 책을 텍스트화 하는 데에 적용하게 된다.

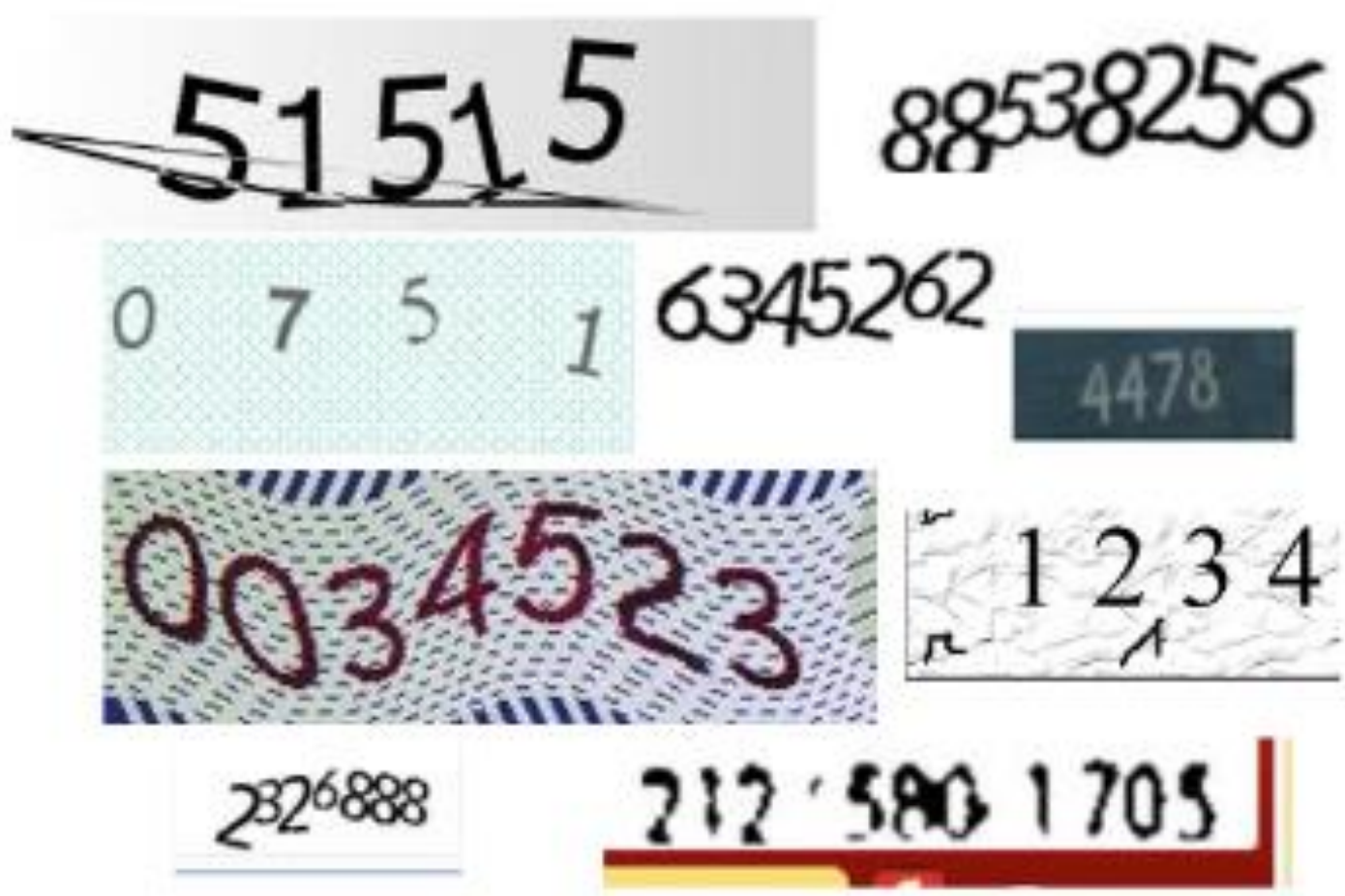
CAPTCHA와 인공지능

CAPTCHA는 인공지능 기술의 한계에 기반을 두고 있다. 예를 들어 왜곡된 텍스트로부터 이미지를 해독해 내는 것은 여전히 쉽지 않다. 그러므로 CAPTCHA는 인공지능 사회에 명확한 과제를 제시한다. 보안을 위해 연구하는 사람들 뿐만 아니라 악의적인 프로그래머들까지도 CAPTCHA를 뚫기 위해 인공지능 분야의 노력을 유도한다. CAPTCHA가 기계에 의해 풀리지 않으면 인간과 컴퓨터를 구분하는 방법이 있는 것이고 CAPTCHA가 풀리면 인공지능의 한계를 해결한 것이다.

이 프로젝트가 해결하고자 하는 문제

홍길동(경찰청 사이버수사대 소속)은 불법적인 방법을 동원한 부동산 투기로 엄청난 부를 축적한 변학도가 그 돈을 모두 비트코인으로 환전하여 보관하고 있다는 것을 알았다. 또 측근을 매수하여 머리가 나쁜 변학도가 비트코인 지갑 주소를 근두운 클라우드에 Text 파일로 저장했다는 사실도 알게되었다. 그 클라우드의 로그인 ID는 변학도가 사용하는 근두운 서비스 이메일이라 알고 있고, 비밀번호는 다른 개발 도구로 알아낼 수 있는데, 근두운 클라우드 로그인 시스템은 아래 그림과 같은 숫자 이미지를 해독하는 CAPTCHA를 로그인에 사용하고 있었다.

이에 홍길동은 CAPTCHA 해독 프로그램을 만들어 해당 시스템 로그인에 성공한 뒤 부정적인 방법으로 부를 축적한 증거를 수집하여 변학도를 검거하고자 한다.



Innovation Academy X Superman Company

프로젝트의 기술적 설명

이 과제는 다음과 같은 단계로 진행된다.

MNIST 데이터셋을 이용하여 하나의 숫자가 표시된 이미지에서 숫자를 인식하는 방법을 익힌다.

위 프로그램을 확장하여 4-8자리 숫자를 인식하는 프로그램을 만든다.

임의의 숫자 이미지 CAPTCHA를 이용하여 컴퓨터에 의한 자동 로그인을 걸러내는 근두운 클라우드의 가상적인 로그인 페이지를 만든다.

위의 로그인 페이지에서 CAPTCHA 이미지를 찾아내서 이미지 내의 숫자를 자동으로 입력하여 로그인을 시도하는 프로그램을 만든다.

결과물에 관한 설명

1. MNIST 데이터셋을 이용하여 숫자를 인식하는 Tutorial을 참고하여 구현한 뒤, 문제를 해결하는 전 과정을 자세히 설명할 수 있는 수준의 이해를 하는 것
2. 위 기본 프로그램을 확장하여 4-8 자리수의 숫자 이미지를 인식하여 출력하는 것.

   주요 Test 데이터는 여러자리의 필기체, 여러 (종류,크기,간격)의 폰트를 여러 기울기로 섞어 출력한 4-8자리의 숫자 이미지. 이 숫자 이미지에 대한 인식률을 85% 이상으로 만들기
3. 가상적 로그인 페이지 만들기. 백그라운드 이미지 위에,

   a. ID(이메일) 입력창, b. 비밀번호(입력하면 * 표시) 입력창, c. 50x160px 크기의 4-8자리의 숫자 임의의 CAPTCHA 이미지가 있으며, d, CAPTCHA 이미지에 대한 답 입력창, e. 백그라운드 이미지나 CAPTCHA 이미지와는 다른 CAPTCHA와 같은 크기의 임의의 이미지가 표시된 가상 로그인 페이지. a,b,c,d,e 다섯 부분이 매번 접속 때마다 임의의 위치에 배치되어야함. c,e 의 이미지는 이미지 형식의 확장자(png)를 가진 파일로, 파일 이름은 8글자의 random 생성된 스트링이어야 함.

   이메일 주소 형식에 맞는 이메일과, 8자리 이상으로 반드시 대소문자, 숫자, 특수문자 조합인 비밀번호가 들어오고, 제시된 CAPTCHA 이미지의 숫자가 맞으면 브라우저에 "OK", 틀리면 'NOT OK"를 출력하는 홈페이지를 로컬 환경에 구성
4. 위 홈페이지에서 CAPTCHA 이미지를 찾아 해독하여 입력하고, 이메일 주소 형식에 맞는 이메일 주소, 비밀번호 규칙에 맞는 임의의 비밀번호를 생성하여 입력하여 "OK", "NOT OK" 비율로 성공 여부를 확인하는 프로그램을 작성
5. 보너스 과제 1
   - 위의 여러 자리수 인식의 예를 차량 번호판 인식으로 확장하여 차량 번호판 인식 프로그램을 작성하여, test set 데이터의 숫자 부분에 대하여 인식률 95% 이상을 달성하는 경우 보너스 포인트 부여
6. 보너스 과제 2
   - 위 보너스 과제에서 차량 번호판의 한글 부분까지를 95%이상의 정확도로 인식하는 경우 보너스 포인트 부여

프로젝트를 구현할 때 제약 조건

1. 4-8 자리수 숫자 인식은 최소 10,000개 이상의 랜덤 이미지를 생성하여 학습을 함.
2. 위 홈페이지 로그인에 사용할 이메일 주소는 '조선실록' 신문사 홈페이지 기사의 byline에서 기자 이메일을 30개 이상 추출하는 프로그램을 작성하여 추출한 뒤 매번 로그인 시도 때 임의의 이메일을 선택하여 사용.
3. 자동 생성된 숫자 CAPTCHA 이미지 내의 숫자는 모두 다른 숫자여야 하며, 가운데 같은 크기, 폰트, 기울기가 같은 모양은 두 개 까지만 허용됨
4. 보너스 문제의 차량 번호판은 "12가3456"와 같은 "숫자2자리 +. 한글1자 + 숫자4자리" 형식의 번호판으로 제한. 번호판 이미지는 실물 번호판의 가로 세로 비율에 맞춰 다양한 픽셀 크기로 흰색, 초록색, 노란색으로 생성하고 인위적으로 잡음을 섞고 (SNR 신호대 잡음비를 xxdB-xxdB 까지), 실제 사진 찍은 결과처럼 사다리꼴 등의 사각형으로 변형한 것 (사각형의 내각이 70도-110도로 제한)을 30% 이상 포함하도록 만들고, 실제 번호판 이미지도 100개 이상을 수집하여 학습에 사용.
5. 과제에서는 머신러닝 부분에서 python 언어와 Tensorflow 프레임워크를 사용.

프로젝트 결과의 검증

1. MNIST Tutorial 구현 내용은 선배 또는 멘토에 의한 코드리뷰와 면접 평가로 확인
2. 여러 숫자 이미지 과제와, 숫자 CAPTCHA 생성 프로그램은 직접 작성하는 것이 테스트 모드에서 사용하며, 결과 검증 모드로 실행할 때에서는 과제의 데이터 파일에 있는 (captcha-gen) 실행 파일을 통해 생성된 00.png ~ 99.png 100개의 이미지를 이용하되, 이름을 이미지를 차례로 읽어 이름을 random하게 바꾸어 적용되도록 함. 그리고 50% 이상의 답을 맞추는 것이 목표임. 맞추는 비율에 따라 포인트가 부여됨
3. 번호판 인식은 본인이 생성, 수집한 데이터로 과제를 수행하며, 검증은 별도의 데이터셋 (license-plate.gz의 이미지)을 이용함

프로젝트 데이터 파일 설명

CAPTCHA-2-T007.tar.gz

doc/CAPTCHA-2-D007.lo.pdf : 이 프로젝트 설명 파일과, 'lo'에 따라 다른 언어 버전.

CAPTCHA-dev-env-image.tar : CAPTCHA 프로젝트 개발 환경이 담긴 Docker Image

captcha-gen : 4-8자리 숫자로 구성된 CAPTCHA 이미지 생성 프로그램 (위 컨테이너 안에서 있고, 그 안에서 실행 가능)

license-plates.gz : 검증에 사용할 차량 번호판 이미지

Innovation Academy X Superman Company

학습 지원 정보

<프로젝트 수행을 위한 학습 리소스>

•[CAPTCHA, Wikipedia](https://en.wikipedia.org/wiki/CAPTCHA) : CAPTCHA에 대한 일반적인 이해 ([CAPTCHA, 한글 위키피디아](https://ko.wikipedia.org/wiki/CAPTCHA)).

•[reCAPTCHA, Wikipedia](https://en.wikipedia.org/wiki/ReCAPTCHA) : reCATPCHA에 관한 이해.

•[Turing Machine, Wikipedia](https://en.wikipedia.org/wiki/Turing_machine) : CAPTCHA가 추구하는 목표를 수학적으로 정의한 Turing Machine에 대한 소개 ([튜링 기계, 한글 위키피디아](https://ko.wikipedia.org/wiki/튜링_기계)).

•[Tensorflow Tutotials](https://www.tensorflow.org/tutorials?hl=ko) : Tensorflow를 초보부터 높은 단계에 이르기 까지 배울 수 있는 곳.

•[모두를 위한 딥러닝](https://www.inflearn.com/course/기본적인-머신러닝-딥러닝-강좌#) : 머신 러닝과 딥러닝에 대해 더 이해하고 본인들의 문제를 Tensorflow를 이용하여 풀 수 있게 도와주는 동영상 강의. 이 강좌는 수학이나 컴퓨터 공학적인 지식이 없어도 어렵지 않게 볼 수 있음.

•[MNIST Tutorial](https://tensorflowkorea.gitbooks.io/tensorflow-kr/content/g3doc/tutorials/mnist/beginners/) : 간단한 컴퓨터 비전 데이터셋인 MNIST를 이용하여 손으로 쓴 글자를 인식하는 과정을 설명.

•[Tensorflow KR 페이스북 그룹](https://www.facebook.com/groups/TensorFlowKR) : tensorflow를 학습하거나, tensorflow, 다른 Machine Learning 도구에 관한 연구, 응용 제작을 하는 분들의 모임. 질의 응답이 매우 활발함.

•[Machine Learning Yearning](https://www.deeplearning.ai/machine-learning-yearning/) : 앤드유 응 교수의 ML 프로젝트를 진행할 때 고려해야 할 우선순위, 성능에 대한 고려 사항 (딥러닝 기술 자체를 다룬 책은 아니라 필독). PDF 버전을 인터넷에서 쉽게 구할 수 있음.

•[AI Hub](https://aihub.or.kr/ai_data) : AI 학습용 데이터가 모여있는 곳.

프로젝트 수행/리뷰/평가/멘토링 Forum

<아래 Forum 사이트에서 이 프로젝트를 수행하는 다른 학생들과 학습 정보를 교환할 수 있으며, 평가,멘토링,코드리뷰를 받을 수 있음>

※ 아래 https://prj.innoaca.kr/forum 은 가상의 도메인임

- 학습 리소스 : https://prj.innoaca.kr/forum/CAPTCHA/2/resource
- 토론과 Q&A : https://prj.innoaca.kr/forum/CAPTCHA/2/forum
- 코드리뷰 : https://prj.innoaca.kr/forum/CAPTCHA/2/codereview
- 동료평가 : https://prj.innoaca.kr/forum/CAPTCHA/2/evaluation
- 멘토 : https://prj.innoaca.kr/forum/CAPTCHA/2/mentor
- 동료목록 : https://prj.innoaca.kr/forum/CAPTCHA/2/peer

Release Note

v007 :

CAPTCHA의 숫자 이미지를 인공지능을 해결하는 Machine Learning 프로젝트 콘텐츠 (예시)

이 예시는 설명을 위해 내용의 많은 부분을 wikipedia에서 가져왔음

아래에 버전 history는 가상적으로 적은 것임 (이전 버전은 없음)

v...

...

v002 :

프로젝트 이미지 데이터 보강으로 프로젝트 완성도 기준 수치 조정

v001 :

CAPTCHA의 숫자 이미지를 인공지능으로 풀어내는 Machine Learning 프로젝트 콘텐츠 설계

프로젝트 정의, 역량정의, 프로젝트 기술, 평가가이드, 데이터 파일 생성

[END-OF-DOCUMENT]

Innovation Academy X Superman Company