Text CAPTCHA 해독 기반 Super Resolution을 활용한 Image Recognition 개선 연구

2017104024 정민혁, 지도 교수: 최진우 교수님

요약

머신 러닝 혹은 딥 러닝은 AI와 Computer Vision(이하CV) 분야에서 다양하게 활용되고 있다. 특히 영상처리 부분에 있어 Image recognition을 통해 객체를 추출해내거나 객체를 인식하여 무엇인지 판별해 내는 많은 연구들이 진행되었다. 따라서 본 연구에서는 CAPTCHA 시스템을 딥 러닝을 통해 학습된 모델로 해독하고자 한 연구를 기반으로 Super Resolution 기술을 도입해 텍스트 이미지 데이터를 가공하여 화질이 개선된 데이터로 학습을 진행하였을 때 기존보다 더 나은 성과를 낼 수 있는지 알아보고자 한다.

1. 서론

1.1. 연구배경

CAPTCHA(Completely Automated Public Turing test to tell Computers and Human Apart)는 사용자를 사람인지 컴퓨터 프로그램인지 구분하기 위해 사용되는 방법의 통칭으로, 다양한 방면으로 활용되지만 주로 로그인이나 회원가입 등에 하나의 절차로 사용된다. 하지만 역설적으로 컴퓨터와 사람을 구분하기 위한 프로그램이 오히려 이를 인공지능 분야의 발전을 낳는 결과를 불러오기도 하였다. Ahn, Blum and Langford의 논문에서 저자들은 CAPTCHA는 AI로 CAPTCHA를 풀지 못하면 컴퓨터와 사람을 구분할 방법이 존재한다는 것이며, 풀었을 경우에는 다른 복잡한 인식 문제에 적용 가능한 AI의 발전을 의미하는 윈-윈 상황을 가져온다고 주장하기도 하였다. AI를 이용하여 CAPTCHA를 풀려는 시도는 과거부터 존재해왔고, 2018년 ACM CCS'18 컨퍼런스에서는 딥러닝 기반으로 상당히 높은 성공률을 가진 공격을 선보인 적도 있다. 머신 러닝을 통해 학습된 모델로 CAPTCHA를 인식하고 해독하는 연구는 충분히 많이 진행된 바가 있다. CNN 구조의 네트워크를 통한 학습이나 적은 데이터 개수로만으로도 학습을 진행하는 연구나, GAN을 통해 임의로 CAPTCHA데이터를 생성하고 그것으로 다시 학습을 진행하는 방법론 등 많은 연구가 선행되었다.

1.2. 연구목표

모델에 대한 연구뿐 아니라 왜곡된 데이터를 다시 이전 형태로 되돌리려는 시도도 존재하며, 본 연구에서는 이러한 관점에서 새로운 시도로 Super Resolution을 접목해보고자 한다. CAPTCHA에 사용되는 이미지들은 주로 컴퓨터가 구분하기 어려운 흐린 이미지나 왜곡된 이미지가 사용된다.

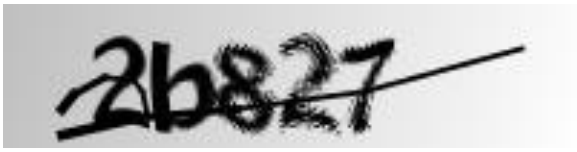

그림 1. CAPTCHA 데이터셋 중 하나

그림 2. CAPTCHA 데이터셋 중 하나

Super Resolution은 저해상도 이미지를 고해상도로 바꾸는 기술이다. 본 연구에서는 CAPTCHA 해독 모델을 학습시키기 전에 전 처리 과정으로 Super Resolution을 거쳐 선명해진 CAPTCHA 데이터셋을 입력 값으로 넣게 될 경우의 모델과 그렇지 않은 경우의 모델의 학습과 테스트 성능을 비교 분석하고자 한다.

2. 관련 기술 및 연구

2.1. CAPTCHA

상기했듯이 CAPTCHA는 Completely Automated Public Turing test to tell Computers and Human Apart의 약어이다. CAPTCHA의 공식 사이트에서는 다음과 같이 CAPTCHA를 소개하고 있다: "A CAPTCHA is a program that protects websites against bots by generating and grading tests that humans can pass but current computer programs cannot." 인간과 컴퓨터를 구분하기 위한 용도라면 형식에 관계없이 다양한 방식이 CAPTCHA에 부합할 수 있다. 가장 많이 볼 수 있는 형태는 이미지에 적힌 텍스트를 읽고 그대로 입력하거나 질문에 답하는 텍스트 CAPTCHA가 있고, 오디오, 이미지를 활용한 방법 등도 존재한다. CAPTCHA의 문제점으로는 간혹 문제가 너무 어려워 사람조차 통과하지 못하는 경우가 발생하거나, 악의를 가진 사람 혹은 그런 목적을 위해 고용된 사람은 구분할 수 없다는 점이 있다. 또한 시간이 흐를수록 클래식한 텍스트 혹은 이미지 CAPTCHA가 머신 러닝과 OCR 프로그램의 발전으로 기계들도 높은 수준의 판독률을 보이게 되면서 본래 용도가 퇴색되고 있다는 점이 있다.

2.1.1. Turing Test

전체 이름에서 알 수 있듯이, CAPTCHA의 튜링 테스트(Turing Test)의 한 종류이다. 튜링 테스트는 1950년 앨런 튜링(Alan Turing)에 의해 제시된 개념으로, 인간이 컴퓨터와 인간을 구분하는데 있어서 기계가 인간이 할 수 있는 것을 할 수 있는지 그 능력을 검증하는 테스트이다. 튜링 테스트의 주체는 인간 평가자로, CAPTCHA는 기계가 컴퓨터와 인간 사용자를 구별한다는 점에서 CAPTCHA는 때로 역 튜링 테스트(Reverse Turing Test)라고 부르기도 한다.

2.1.2. reCAPTCHA

reCAPTCHA는 CAPTCHA의 한 종류로, Luis von Ahn, David Abraham, Manuel Blum, Michael Crawford, Ben Maurer, Colin McMillen, and Edison Tan에 의해 개발되었고, 2009년부터 구글에 소유되어 무료로 사용가능한 서비스이다. V3까지 출시한 상태이며, 일반적으로 웹사이트에서 많이 사용되는 경우는 V2이다. reCAPTCHA V2는 체크박스를 클릭하여 봇이 아닌 사람임을 인증 받고, 몇몇 경우에는 흔히 아는 이미지 CAPTCHA를 추가적으로 진행하여 인증을 받는 방식이다.

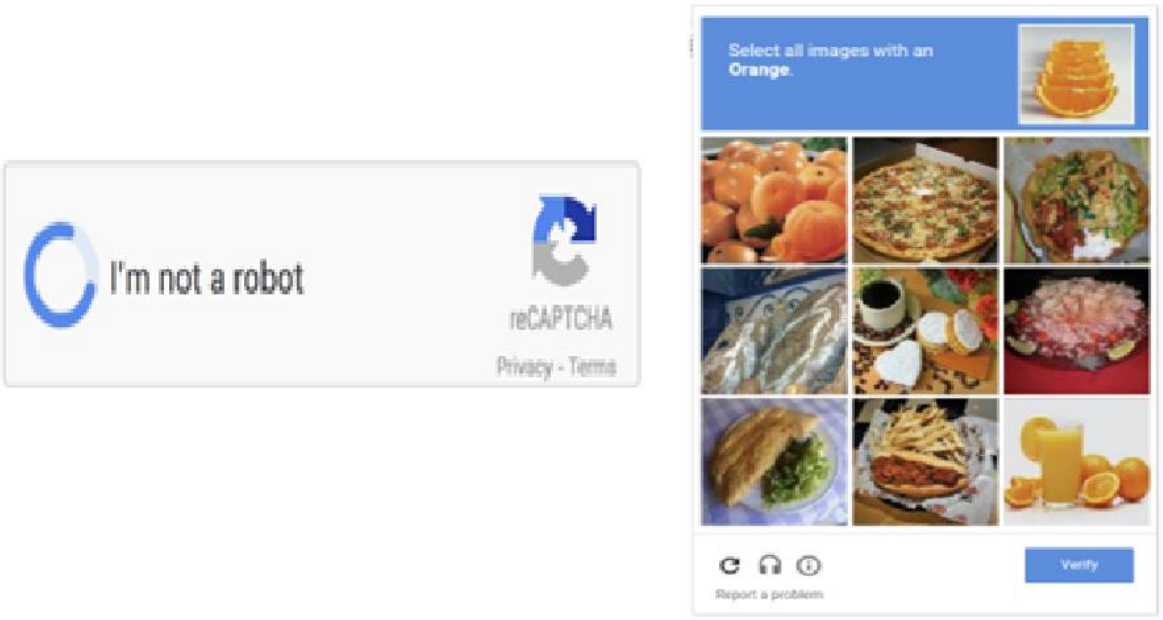

그림3. 구글 reCAPTCHA (이미지 출처: https://support.google.com/recaptcha/?hl=en)

reCAPTCHA는 고문서들이나 OCR 프로그램으로 제대로 인식이 되지 않는 텍스트를 판독하고 입력하기 위해 인간의 힘을 빌려 사용자를 봇이 아님을 판독함과 고문서의 디지털화 작업까지 같이 이루는 프로젝트를 진행하기도 하였다. 또한 이미지 CAPTCHA를 통하여 다량의 학습데이터를 축적하는 일거양득을 이루기도 하였다. reCAPTCHA V3는 인증과정을 거치면서 발생하는 사용자의 불편함을 없애기 위해 이러한 과정을 없애고, 사용자의 행동에 점수를 매겨 이에 따른 대응을 상황에 맞게 하는 프로그램이 자동으로 돌아가게 바뀌었다.

2.2. OCR

OCR(Optical Character Recognition)은 광학 문자 인식을 뜻하며, 텍스트 이미지에서 수정, 검색, 저장, 표기, TTS(text-to-speech)등의 목적을 위해 텍스트를 추출하여 전환하는 기술을 뜻한다. 패턴 인식, 인공지능, 컴퓨터 비전을 아우르는 연구 분야이며, 머신 러닝을 동반하기도 한다. OCR은 전처리, 텍스트 인식(패턴 매칭+특징 추출), 후처리 과정으로 이루어져 있다. 상기한 목적 외에도 CAPTCHA의 안정성을 테스트하기 위해 OCR이 사용되기도 한다.

2.3. 머신 러닝

머신 러닝은 컴퓨터과학의 한 분야로 경험을 통해 학습 또는 발전하는 알고리즘을 연구한다. AI연구의 한 분야로 간주되며, 모델을 구축하여 샘플 데이터를 통해 결과를 얻고, 이 결과를 이용하여 모델을 다시 수정하는 과정을 반복하면서 입력 값에 대해 원하는 결과를 출력하도록 수정하는 과정을 거친다. Tom M. Mitchell은 "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." 라고 컴퓨터의 학습을 정의하였다. 어떤 경험이 작업의 평가를 더 좋게 만들면 그 경험을 통해 프로그램이 학습했다고 정의한 것이다. 이는 즉 데이터가 많으면 더 많은 경험을 통해 학습할 기회가 있음을 뜻하고, 빅데이터의 중요성을 불러오기도 하였다. 인간이 방법이나 규칙을 알고 있는 현상이나 작업에 대해서는 프로그램을 설계하는 것이 가능하지만, 규모가 지나치게 방대하거나 인간 역시 제대로 이해하지 못하는 영역은 프로그래밍이 어려우며, 이런 경우의 알고리즘을 구축하는데 머신 러닝이 유용하게 사용된다. 모델을 구축할 때 모델을 이루는 네트워크 레이어의 개수가 많은 경우(일반적으로 3개 이상) 이를 딥 러닝이라고 부른다. 자주 사용되는 네트워크 구조로는 CNN이 있다. 본 연구에서도 CNN을 활용한 모델을 포함할 예정이다.

2.3.1. CNN

CNN은 딥 러닝에 주로 사용되는 네트워크 구조의 한 종류로, 커널의 공유되는 weight를 통한 convolution 연산을 통해 모델에 있는 모든 노드들이 완전 연결되어 있는 것이 아닌 부분연결 구조를 가져서 모델의 복잡도가 낮은 모델이다. Convolution연산은 본래 영상처리나 신호처리 분야에서 많이 사용되었고, 따라서 CNN역시 Image Classification, Image Segmentation, Video Recognition등과 같은 분야에 많이 사용된다.

2.4. Super Resolution

Super Resolution은 저해상도 이미지를 고해상도로 변형 또는 복원시키는 컴퓨터 비전의 연구 분야이다. Papers With Code에서는 Super Resolution을 “Augmenting and increasing the resolution of an image using classic and advanced super-resolution techniques” 와 같이 정의하고 있다. 딥 러닝 이전에는 Bicubic Interpolation 혹은 Linear Mapping 등의 방식들이 주로 사용되어왔으나, 기존 전통적인 복원 기법보다 딥 러닝에 기반한 기법들이 뛰어난 성능을 보이면서 딥 러닝에 기반한 연구가 계속되고 있는 추세이다. 잘 알려진 네트워크로는 SRCNN, VDSR, SRGAN등이 있다. 본 연구에서는 SRGAN을 사용하여 CAPTCHA이미지의 upscaling을 진행할 예정이다.

2.4.1. GAN

GAN(Generative Adversarial Network)은 2014년 Ian Goodfellow에 의해 최초로 제안된 머신 러닝의 모델의 한 종류로, 생성기와 분별기 두 개의 네트워크가 서로 경쟁하면서 서로의 학습을 유도

하는 방식으로 구현된 비지도 학습에 유용한 모델이다. GAN은 생성기가 분별기가 구분하기 어려울 정도의 샘플을 만들어 내는 것을 목표로 한다. 최초의 GAN은 DMLP를 사용하여 구현되었고, 이후 CNN을 사용한 DCGAN등이 나왔다. SRGAN에서는 저해상도 이미지를 고해상도로 만들면서 upscaling을 할 때 GAN 모델을 통해 빈 자리를 생성한다.

3. 프로젝트

3.1. 프로젝트 진행방법

프로젝트는 다음의 순서로 진행할 예정이다.

1. CAPTCHA 해독을 학습하기 위한 간단한 모델을 직접 구축한다.
2. 구축한 모델로 CAPTCHA 데이터셋을 학습 및 테스트한다.
3. CAPTCHA cracker 라이브러리를 사용한 모델을 통해 동일 데이터셋을 학습 및 테스트한다.
4. SRGAN을 통해 CAPTCHA 데이터셋을 고해상도로 만든 결과물을 얻는다.
5. 2-3의 과정을 고해상도 데이터셋을 통해 진행한다.
6. 2-5의 각각의 학습 및 테스트 결과를 비교한다.
7. 위의 과정을 다른 데이터셋으로 반복한다.

3.2. 사용데이터

훈련 및 테스트에 사용할 모델은 두 종류이다. 두 데이터셋 모두 Kaggle에서 무료로 이용 가능하다. 1번 데이터셋은 약간의 왜곡이 가해진 숫자와 알파벳이 섞인 5글자 흑백 CAPTCHA 이미지들로 구성되어 있고, 2번 데이터셋은 색깔과 배치가 랜덤인 숫자와 영어가 섞인 10글자 CAPTCHA 이미지들로 구성되어 있다.

1. https://www.kaggle.com/datasets/fournierp/captcha-version-2-images?resource=download (original source: https://www.researchgate.net/publication/248380891_captcha_dataset)
2. https://www.kaggle.com/datasets/aadhavvignesh/captcha-images

3.3. 사용모델

Super Resolution을 적용하기 위해서는 상기한 SRGAN을 사용한다. CAPTCHA 이미지 학습을 위한 모델로는 CNN 기반 모델을 사용한다.

4. 향후 일정

10/1 ~ 10/15: 학습 모델 구축 및 작동 테스트

10/16 ~ 10/23: 데이터셋에 Super Resolution을 적용하여 개선된 데이터셋 획득 및 학습

10/23 ~ 10/25: 중간 보고서 작성

10/26 ~ 11/7: 결과 비교 및 개선점 탐색

11/8 ~ 11/25: 최종 보고서 작성 및 발표 준비

5. 결론 및 기대효과

[그림1]과 [그림2]에서 볼 수 있듯이 샘플 데이터들은 노이즈가 심한 저해상도 이미지이다. 따라서 Super Resolution을 통해서 해상도를 개선했을 때 선명한 이미지로부터 더 많은 특징을 잘 추출해 내어 더 빠른 속도로 학습이 진행되리라고 예상하며, 개선된 학습 효율이 기대된다. 또한 만약 성능 개선이 관측될 경우 이를 일반적인 이미지 학습에도 충분히 확장 적용시킬 여지가 있다고 생각한다. 더하여, Super Resolution을 거친 방법이 학습 결과가 더 좋을 경우, 이는 기존 CAPTCHA 시스템의 변화가 더더욱 요구됨을 의미한다. 구글의 reCAPTCHA V3가유저가 테스트를 받는 형식에서 벗어나 유저 행동에 점수를 매기는 것처럼 CAPTCHA 방식 자체에 변화를 주거나 사람과 컴퓨터를 구별하기 위한 완전히 새로운 방법론이 요구될 것이다.

6. 참고문헌

1. von Ahn, Luis; Blum, Manuel; Hopper, Nicholas J.; Langford, John (May 2003). "CAPTCHA: Using Hard AI Problems for Security". Advances in Cryptology — EUROCRYPT 2003. EUROCRYPT 2003: International Conference on the Theory and Applications of Cryptographic Techniques. Lecture Notes in Computer Science. Vol. 2656. pp. 294–311.

2. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4681-4690).

3. Mitchell, T. M., & Mitchell, T. M. (1997). *Machine learning* (Vol. 1, No. 9). New York: McGraw-hill.

4. 서유림, 강석주 (2020). "특집 최신 미디어와 인공지능 - 딥러닝 기반 Super Resolution 기술의 현황 및 최신 동향". 방송과 미디어 제25권 2호" pp. 7-15.

5. 오일석, (2017). 기계학습 pp. 23-25, 194-200, 225-229.

6. Wikipedia, "튜링테스트" (2022/08/29) https://ko.wikipedia.org/wiki/%ED%8A%9C%EB%A7%81_%ED%85%8C%EC%8A%A4%ED%8A%B8

7. Wikipedia, "reCAPTCHA" (2022/09/14), https://en.wikipedia.org/wiki/ReCAPTCHA

8. Wikipedia, "Optical character recognition" (2022/09/24), https://en.wikipedia.org/wiki/Optical_character_recognition

9. Wikipedia, "Machine Learning"(2022/09/28), https://en.wikipedia.org/wiki/Machine_learning

10. Wikipedia, "Convolutional neural network" (2022/09/21), https://en.wikipedia.org/wiki/Convolutional_neural_network

11. Wikipedia, "CAPTCHA", (2022/06/20), https://ko.wikipedia.org/wiki/CAPTCHA

12. Wikipedia, "CAPTCHA", (2022/09/17) https://en.wikipedia.org/wiki/CAPTCHA

13. Wikipedia, "Generative adversarial network" (2022/09/25), https://en.wikipedia.org/wiki/Generative_adversarial_network

14. CAPTCHA official website, http://www.captcha.net/

15. Google Help, "reCAPTCHA Help", https://support.google.com/recaptcha/?hl=en

16. Google reCAPTCHA, https://www.google.com/recaptcha/about/

17. AWS, "광학 문자 인식(OCR)이란 무엇인가?", https://aws.amazon.com/ko/what-is/ocr/

18. Papers with Code, Computer Vision - "Image Super-Resolution", https://paperswithcode.com/task/image-super-resolution