# 인공지능 데이터 셋을 위한 이미지 웹 크롤링 프로그램

박은수 양영준 전지호 *류은석

가천대학교

dmsehf804@gc.gachon.ac.kr *esryu@gachon.ac.kr

# Image Web Crawling Program for Artificial Intelligent Data Set

Park, EunSoo Yang, YoungJun Jeon, JiHo *Ryu, Eun-Seok

Gachon University

## 요약

최근 인공지능에 대한 연구가 많아지면서 데이터 셋에 대한 연구 또한 많아지고 있는 추세이다. 하지만, 기존 이미지 웹 크롤러들은 사용자가 원하는 이미지뿐만 아니라 내용상 맞지 않는 이미지들도 많이 포함되어 있다는 문제점을 가진다. 본 캡스톤 디자인은 최근 연구주제로 떠오르는 인공지능을 이용한 정확도가 높은 이미지 웹 크롤링 프로그램을 개발 내용을 설명한다. 프로그램 구동과정은 (1) 크롬 드라이버를 통하여 사용자가 원하는 키워드를 입력 받은 후 해당 URL로 접근하여 이미지들을 모두 다운로드 받는다. (2) 그 후 객체 감지 프로그램을 사용하여 사용자가 원하는 이미지를 잘라낸다. (3) 그리고 추가적인 이미지 필터 기능을 선택적으로 적용한다. 해당 기능을 통하여 사용자가 인공지능을 학습할 때 원하는 필터링을 적용할 수 있다. 결론적으로, 본 캡스톤 디자인은 위의 기능들을 이용함으로써 사용자의 이미지 객체 요구에 높은 정확도를 가지는 이미지 크롤러 개발 과정 및 결과를 설명한다.

## 1. 서론

인공지능은 인간의 지능적인 능력과 기능들을 기계가 대신하도록 하는 것을 목적으로 하는 지능과 그 응용을 다루는 학문이다. 현재는 인공지능이 자동차 산업, 언어 처리 관련 산업, 의료 산업, 금융 산업 등에 사용되고 있다[1]. 인공지능은 2010년대를 전후로 클라우드 컴퓨팅 및 빅데이터의 등장, 컴퓨팅 파워의 개선 및 네트워크의 활성화, 딥 러닝 등 알고리즘 발전으로 기술력이 급성장하며 다시금 각광을 받기 시작하였다[2].

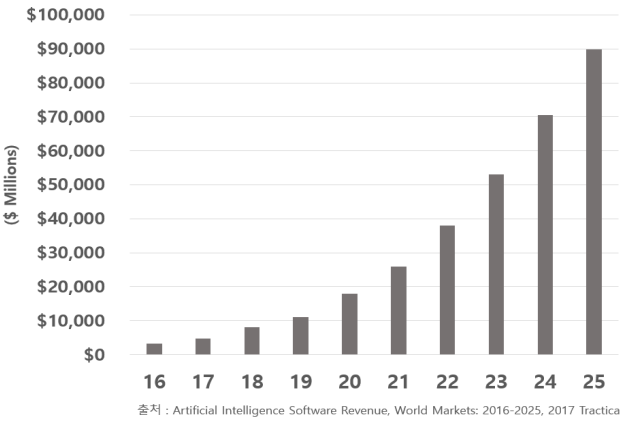


그림 1. 전세계 인공지능 소프트웨어 시장 전망

그림 1과 같이 전 세계적 인공지능 소프트웨어 시장은 2016년의 32억 달러부터 시작하여 2025년에는 898억 달러까지 전망을 보이고 있다.

인공지능 특히, 합성곱 신경망의 경우 학습을 하기 위해서 대량의 이미지 데이터가 필요하다. 일반 사용자들이 데이터를 얻는 경로는 ImageNet[3], 이미지 웹 크롤링, 특정 대학에서 배포하는 데이터 셋 등이 있다. 그러나 해당 경로들은 대량의 데이터를 얻을 수 있으나, 그림 2와 같이 노이즈 즉, 원하는 객체가 포함되지 않은 사진들이 많이 있을 수 있다. 본 캡스톤 디자인 보고서에선 사용자가 원하는 객체에 대한 최소한의 노이즈가 적용되어 있는 데이터 셋을 얻을 수 있는 프로그램을 설계 및 구현한다.

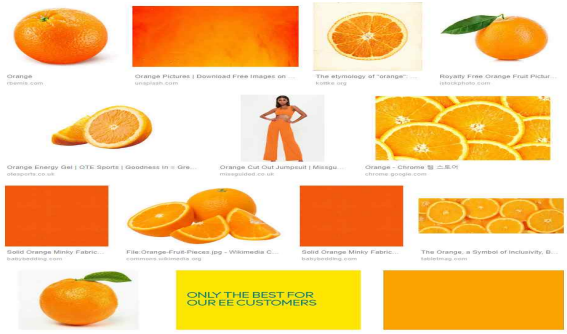


그림 2. Orange를 검색할 경우 나오는 노이즈 이미지

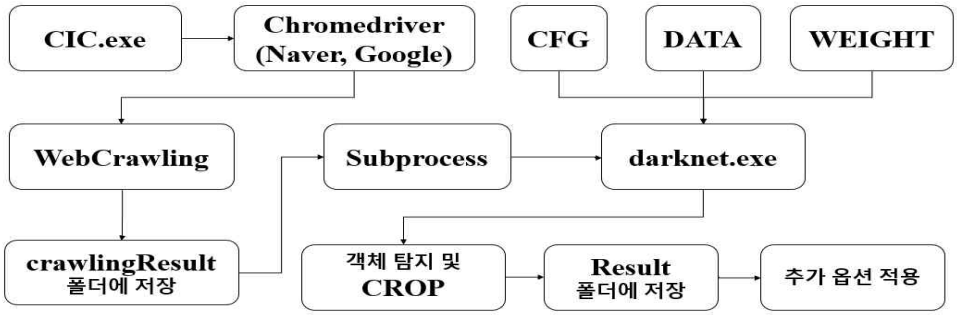


그림 3. 프로그램의 전체적인 흐름

## 2. 제안하는 이미지 크롤러의 설계 및 구현

본 프로그램의 전체적은 흐름은 그림 3과 같다. 객체 탐지 프로그램은 타 프로그램들에 비해 속도가 월등히 빠르면서 mAP도 높은 darknet의 YOLO V2 for Windows(이하 YOLO)[4]를 사용하였다. 이미지 키워드는 YOLO를 학습할 때 사용한 COCO dataset[5]의 클래스들을 사용하였다. 클래스는 총 80개며 UI에서 오른쪽에 배치하였다. 스크롤링을 통하여 검색할 수 있는 키워드들을 볼 수 있다. 사용자는 원하는 키워드를 클래스 리스트에서 찾아 검색을 하여 이미지 웹 크롤링을 한다. 이미지 웹 크롤링은 크롬 드라이버를 사용하여 구글, 네이버 이미지 검색을 하여 이미지 URL을 받아 다운로드 하였다. 다운 받은 이미지들은 crawlingResult 폴더에 저장된다. 사용자가 검색한 키워드는 동일 폴더의 label.txt에 텍스트로 저장이 되며 이 후 YOLO가 실행된다. YOLO는 크롤링 한 이미지들에서 label.txt에 입력되어 있는 키워드만을 탐지하며 크롭하여 result 폴더에 저장한다. 저장된 이미지들은 프로그램의 기능인 필터와 리사이즈 옵션을 적용할 수 있다. 필터는 OpenCV에서 제공하는 함수들로 제작하였으며 적용하면 result_필터명의 폴더에 저장된다. 리사이즈 옵션은 사용자가 원하는 이미지 크기(가로, 세로)를 입력한 후 버튼을 클릭하면 resize_가로_세로 폴더로 이미지들이 리사이징 되어 저장된다.

본 캡스톤 디자인의 작품은 Python 3.5.2 버전과 OpenCV 2.4.9버전을 사용하였다. 배포를 위해 pyinstaller를 사용하여 main 파일을 실행 파일로 만들었다. YOLO는 C언어로 제작되어 Visual Studio 2015 버전에서 개발 및 수정을 하였고 CUDA 8.0이 필요하다. 또한 두 가지 프로그램을 Python의 subprocess 기능을 사용하여 두 프로그램을 연결하였다.

개발 및 실험을 한 PC의 사양은 Windows 10 Home 64비트, i7-7700 CPU 3.60GHz, 8GB 메모리, GeForce GTX 1080이다.

## 3. 구현 결과

구현된 프로그램의 UI는 그림 4와 같으며 검색란, 필터 옵션, 리사이즈 옵션, 검색키워드로 구분되어 있으며, 옵션들을 적용할 때 간단한 메시지창이 팝업된다. apple 키워드를 입력하고 검색한 후 그림 5와 같이 crawlingResult 폴더에 노이즈가 많은 이미지들이 crawling 되고 그 후 객체 탐지 과정을 거쳐 result에 사과이미지만 저장되는 것이 보인다. 마지막으로 그림 6과 같이 추가옵션을 적용하여 해당 폴더에 필터, 리사이즈가 적용되는 것이 보여진다.


그림 4. 프로그램 UI

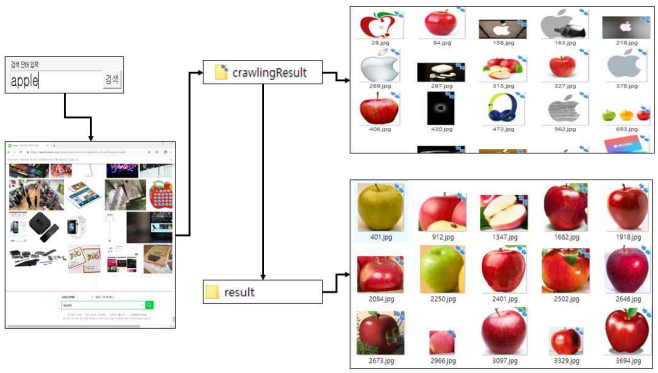


그림 5. 키워드 검색 후 산출물


그림 6. 추가 옵션 적용 후 산출물

## 4. 기대 효과 및 결론

본 프로그램은 단순한 이미지 크롤링 프로그램을 넘어서서 인공지능을 사용하여 사용자가 원하는 객체만을 탐지하고 크롭하여 노이즈가 일반적인 프로그램들에 비해 적은 데이터 셋을 얻을 수 있는 프

로그램이다. 또한 필터링 및 리사이즈를 통하여 앞으로도 많은 연구가 있을 인공지능 즉, 합성곱 신경망의 이미지 데이터 셋을 모으는 데에 많은 기여를 할 것이며, 단순히 프로그램을 다운받고 압축을 해제한 후 실행파일을 실행만하면 누구나 사용할 수 있다는 것이 장점이다.

## Acknowledgement

본 연구는 한국전력공사의 2016년 선정 기초연구개발과제 연구비에 의해 지원되었음 (과제번호 : R17XA05-68)

## 5. 참고문헌

[1] 정동규. (2017). 인공지능 기술과 주요 적용 산업 동향. 한국정보기술학회지, 15(2), 21-28.
[2] 원동규, 이상필. (2016). 인공지능과 제4차 산업혁명의 함의. ie 매거진, 23(2), 13-22.
[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248-255.
[4] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788
[5] Lin TY. et al. (2014) Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham