**프로젝트 주제: 웹툰 글자 이미지 텍스트로 변환하기**

**OCR기술을 활용하여**

**퓨처사이언테크인턴십**

**김 한**

|  |
| --- |
| **<목차>**   1. **OCR기술이란?** 2. **OCR기술의 활용** 3. **OCR의 기술적 원리** 4. **프로젝트 고려사항** 5. **활용가능한 오픈소스** 6. **관련 기술 공부** 7. **프로젝트 설계 계획** |

**1.OCR 기술이란?**

광학 문자 인식기(Optical Character Recognition)는 **사람이 쓰거나 기계로 인쇄한 글자 이미지**를 **텍스트로 변환**하여 **인식하는 기술**을 말한다.

Object를 인식하는 것에서 좀 더 확장하여 text를 인식하는 것이 OCR 기술의 핵심이므로 object detection의 활용이라고 할 수 있음.

* 따라서 **object Detection에 대한 공부가 선행**되어야 하며, 그보다 **앞서 image recognition & localization 을 먼저 공부**해야 할 것으로 보임.

**2.OCR 기술의 활용**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **차량번호판의 텍스트 이미지를 인식.** |  | **네이버의 스마트렌즈 – 카메라로 특정 물체의 텍스트 이미지를 인식하고 번역함.** |

**3.OCR의 기술적 원리**



**Text localization** : 이미지 속에서 **text의 위치(영역)**를 찾아내는 관련 기술.

**Text recognition** : localization을 통해 찾아낸 text 필드에서 **글자를 읽어내는(글자 각각을)** 기술.

유사 관련기술 : Object Detection

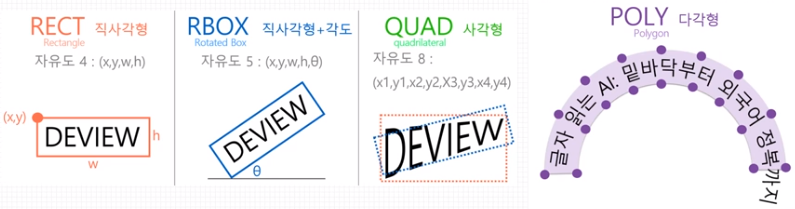


OCR의 주요 이슈

Object detection과 달리 detection box의 **종횡비가 일정하지 않음.**

**글자의 크기**, **기울어짐의 정도**, **글씨체 등의 다양함** 때문에 구별이 쉽지 않다는 이슈가 존재함.

**어떻게 Text Localization 을 할 것인가?**

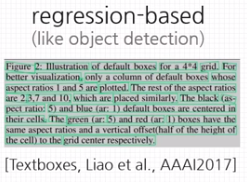


이와 같이 기본적인 직사각형, 기울어진 경우, text의 왜곡이 있는 경우, 보다 둥글게 잡힌 경우가 존재함.

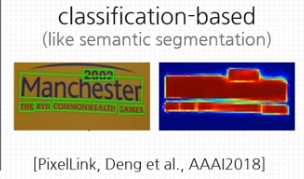
각각의 모델의 자유도를 늘려가며 다양한 text의 detection이 가능함.

이를 위한 방법으로 3가지가 존재한다.

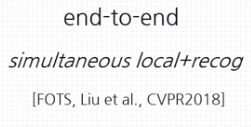
1. Regression based 방식임. Object detection의 기본원리에 박스의 종횡비를 수정하며 보안한 방식임. 이미지를 거치면 CNN을 글자영역표현 값을 바로 예측하는 방식.



1. segmentation-based 방식임. CNN을 거친 출력 값이 화소 정보단위와 간단한 규칙으로 글자 영역 표현 값들을 예측하는 방식. 출력 값으로 글자영역에 속할 영역들을 나오게 됨.



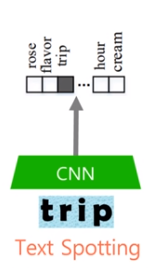
1. End-to-end 방식임. Localization 과 recognition 방식을 한번에 해내는 방식임. 한번에 접근함으로써 2가지 모델의 방식에 긍정적영향을 주는 방식.



최근에 가장 각광받는 방식은 segmentation방식인 classification-based 이다. 그 이유는 training과 구현이 가장 간단한 편이기 때문이다.

**어떻게 Text recognition 을 할 것인가?**

1. Text spotting 방식 : 글자를 인식하는 일을 단어를 분류하는 classification 문제로 접근하는 방법론. 바로 단어와 매칭하는 방식인데 단어가 너무 많아서 이를 고려하기 어려움.

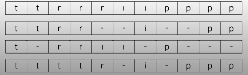


1. CTC 방식 : 규칙기반 후처리 방식. 이미지가 들어와 CNN을 거치면 특징이 추출되고 그 특징을 RNN을 통과해서 고정길이의 sequence 정보를 뽑아내게 됨. 이를 CTC라는 후처리 방식을 통해 접근하는 방식.



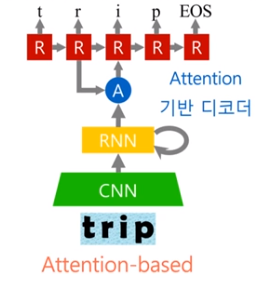
구체적으로 CTC 방식은 무엇인가?

Detection된 text를 sliding window를 하며 각각의 글자를 하나씩 인식하는 방식.



이런 식으로 sliding window를 통해 고정길이의 시퀀스를 얻어내고 이를 **적절한 후처리**를 통해 trip이라는 단어를 뽑아내는 방식임. [CRNN & GCRNN 방식을 활용]

1. Attention-based 방식 : 가변길이 시퀀스 방식으로 마찬가지로 CNN과 RNN을 거치고 고정길이의 sequence 정보를 뽑아내게 됨. 이 sequence를 attention 기반 Decoder가 잘 해석하여 정답을 만들어내는 접근 방식.



Attention-based 방식은 구체적으로 무엇인가? Attention은 RNN의 현재 스텝의 결과를 뽑기 위해서 이전 스텝의 정보들 중 어떤 정보를 더 중점적으로 볼지에 대한 가중치를 말함. [RARE & FAN 방식을 활용]

요즘추세는 **attention-based 방식이 각광**을 받고있고, **속도가 느린** 반면 **높은 정확도**를 가짐. 대용량의 경우에는 속도가 빠른 CTC 방식이 어울리다고 할 수 있음.

**4.프로젝트 고려사항**

1. 웹툰의 경우는 **사용자가 읽어야 하므로 가독성이 떨어지는 글씨일 확률이 적으므로** text recognition 단계에서 보다 가볍게 구현이 가능할 것으로 보임.
2. **세로로 글씨를 쓰는 경우도** 있으므로 이 역시 고려를 해야 할 것으로 보임.
3. **말풍선을 먼저 detection**을 해내면 text localization 단계에서 **높은 정확도**를 보일 것으로 보임. 하지만 웹툰 내에 **말풍선 이외에 text 정보가 굉장히 많으므로 이 역시 detection 하는 것이 중요**하다고 고려 됨.

* 따라서 3번의 경우엔 phase 1로 말풍선 내의 text를 먼저 처리하고 이를 성공 할 경우엔 phase 2로 전체 이미지 내에서의 text 정보를 찾아내는 것을 목표로 하면 좋을 것으로 고려 됨.

1. 무엇보다 tesseract를 오픈소스로 보통 활용하는데 **한글에 대한 정확도가 굉장히 떨어짐**. 따라서 **조합가능한 모든 한글 글자를 이미지화하고 이를 학습 데이터로 활용**하는 방식을 사용.

* 관련논문 : Deep Learning을 이용한 한글 OCR 정확도 향상에 대한 연구<http://www.ndsl.kr/ndsl/commons/util/ndslOriginalView.do?dbt=CFKO&cn=CFKO201821464987120&oCn=NPAP12688214&pageCode=PG11&journal=NPRO00377584>

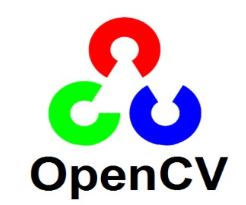
**5.활용가능한 오픈소스**

왜 오픈소스 이어야 하는가?



Abbyy와 adobe 같은 OCR의 경우 높은 정확도를 자랑하지만 유료이므로 본 프로젝트의 적용에 문제가 있다. 따라서 오픈소스를 활용하여 프로젝트를 진행한다.

어떤 오픈소스가 존재하는가?

1.  open CV는 이미지 및 영상 처리부분에서 가장 권위가 있는 오픈소스이다. C++기반으로 작성 되었으며 자바와 파이썬 등의 언어 역시 제공한다. Open CV는 OCR에서 이미지 전처리와 Text Localization에 사용 가능하다.
2.  tesseract OCR은 OCR 부분의 오픈소스 강자로 2018년엔 딥러닝 학습 기반이 내장되어 있는 강력한 툴이다. V4가 배포되어 있으며 LSTM(RNN의 일종)기반으로 개발되어 인식 정확도가 크게 높아졌다.

따라서 본 프로젝트는 **open CV를 활용하여 전처리를 하는 text localization을 진행**하고 **Tesseract OCR의 딥러닝 학습기능을 이용하여 text localization과 text recognition의 성능 역시 향상** 시킴으로써 웹툰 text 이미지를 **정확하게 text로 translation하는 것을 목표**로 한다.

**6.관련 기술 공부**

Machine Learning 과 Deep Learning

의 관계이다. Machine Learning의 한 방법인 Neural Network를 Deep하게 여러 층으로 쌓아서 보다 훌륭한 퍼포먼스를 내는 것이 Deep Learning 이라고 할 수 있음.

Deep Learning 모델의 소개

**Fully-connected Neural Network** : 신경망에서 퍼셉트론을 다중으로 연결한 방식을 말함.

**Convolution Neural Network** : FC Net과 달리 가로세로높이의 **큐브 형태로 퍼셉트론을 정의**하고 **Hidden Layer를 Filter로 사용하는 방식**. Filter에 속한 각각의 퍼셉트론은 모두 다른 weight를 가지고 있으며, 기존의 수학적으로 Filter의 weight 값을 설정한 것과 달리 Deep Learning에서의 **Filter는 수 많은 이미지 Data의 학습을 통해 최적의 Filter 의 weight 값을 찾아가는 것**을 목표로 한다.

RCNN(Region with CNN): RCNN은 Object Detection 문제를 풀기 위함이다. 어떤 이미지가 주어지면, 그 이미지에 있는 모든 object들에 대해 bounding box를 그리는 것이 최종 목적이다. 따라서 Region Proposal 과정과 classification 과정으로 나뉜다.

**Recurrent Neural Network :** 시퀀스 **데이터의 길이 가변성과 선후 관계를 학습할 수 있는 방식의 Network**로써 FC Net의 연결구조에서 **이전 Hidden Layer의 벡터 값을 같이 전달하여 해당 Hidden Layer의 가중시키는 방식**이다. 이러한 방식을 채택하기 때문에 선후관계 정보가 중요한 시퀀스 데이터에 대해서 잘 작동하며 오늘날 자연어 처리에 특히 적용되고 있다.

**Deep Learning의 강점**

기존 Machine Learning의 경우 feature를 정의하고 추출하는 것이 성패에 지대한 영향을 미친다. 그러나 Deep Learning에선 **요인 표현 학습(Feature Representation Learning)이라 하여 원본 데이터로부터 최적의 성능을 발휘하는데 사용될 수 있는 Feature들을 스스로 학습**하고 이에 맞는 Weight를 효과적으로 찾을 수 있다.

**Image Classification?** 주어진 이미지 안에 어느 특정한 **클래스에 해당하는 사물이 포함되어 있는지 여부를 분류**하는 모델을 만드는 것을 주요 목표로 한다. 여기서 클래스란 분류 대상이 되는 카테고리 하나하나를 지칭한다. 신뢰도 점수를 기반으로 들어있는 이미지의 클래스가 무엇인지를 구별해낸다.

**Image localization(Detection)?** 주어진 **이미지 안에 어느 특정한 클래스에 해당하는 사물이 있다면** **어느 위치에 포함**되어 있는지 **박스형태(bounding Box)**로 검출하는 모델을 만드는 것을 목표로 하는 분야. 이는 특정 클래스의 사물이 포함되어 있는지 여부만을 분류하는 Classification 문제의 목표에서 한 층 더 나아간 것이라고 할 수 있다.

Bounding Box의 정의: 전체 이미지 상에서 **박스의 좌측 상단의 좌표 (x1,y1)과 우측 하단의 좌표 (x2,y2)를 결정**해야 한다. 이와 더불어 제시한 Bounding Box 안에 포함된 사물에 대한 각 클래스 별 신뢰도 점수도 제시함으로써 이것이 무엇인지에 대한 결과도 제시해야한다.

**Image recognition?** 기계로 하여금 주어진 이미지 상에 포함되어 있는 **대상이 무엇인지(Image Classification), 또한 어느 위치에 있는지(image Localization)** 등을 파악하도록 하는 것을 주된 목표로 한다. **즉, recognition = localization + classification**

**Object detection?** Image라는 광범위한 범위에서 포함되는 Image 내에서의 Object(물체,객체)를 찾는 좁은 범위의 분야라고 할 수 있다. 마찬가지로 Recognition은 Detection + Classification이라 할 수 있음.

**Tesseract API의 내부 Logic?**

**Tesseract 에서 제공하는 한글 OCR에 대한 문제점 분석.**

1. 다른 폰트에 대해서 구별이 가능한가? **띄어쓰기도 엉망일때가 있고 기울어진 글씨체나 모든 폰트에 대해서 인식을 하는 것이 아님**. 학습이 필요해 보인다.

|  |  |
| --- | --- |
|  |  |