

국립대학육성사업 자연과학대학 기초보호학문 R&D 인턴십 프로그램 결과보고서

연구실명	수학과 과학계산연구실
작성자(학부생)	학번: 2017010715 (4학년) 성명: 허지혜
교육조교	성명: 이승규 교수님
교육주제	딥러닝을 이용한 Tesseract OCR
운영기간	2020. 10. ~ 2021. 01.

차례

1. 월별 중점 운영 내용
2. 2020년 10월 활동 보고서
3. 2020년 11월 활동 보고서
4. 2020년 12월 활동 보고서
5. 2021년 01월 활동 보고서
6. 결론 & 느낀 점

1. 월별 중점 운영 내용

주제 : Deep Learning을 이용한 Tesseract OCR

2020년 10월	Tesseract 개념을 살펴보고 실제 사용 예시들을 공부하였다. Tesseract에 사용되는 딥러닝의 개념과 머신 러닝, 딥 러닝의 차이, 딥 러닝의 역사 등등을 살펴보았다.
2020년 11월	딥 러닝에 대해 필요한 수학적 지식들을 먼저 공부하고 Tesseract에 들어가는 딥 러닝 모델들에 대하여 인지하였다. Tesseract에 쓰이는 딥러닝 모델인 CNN에 대해 공부하기 전에 가장 기본적으로 쓰이는 DNN 구조에 대해서 먼저 실습을 한 후, CNN에 대하여 공부를 하였다.
2020년 12월	Tesseract에 들어가는 딥러닝 모델인 RNN과 더 나아가 LSTM까지 학습하고 실습하였으며, 들어가는 손실함수는 CTC LOSS에 대하여 공부를 하였다.
2021년 01월	위에서 했던 모델들에 대해서 각각의 실습을 해보았다. Tesseract를 담고 있는 OCR에 대한 공부를 하고 구글에서 지원하는 Tesseract로 실습을 시도하였다.

2. 2020년 10월 활동 보고서

Chapter 1. 딥러닝 시작

1.1. 딥러닝과 머신러닝의 관계

1.2. 머신러닝 의미

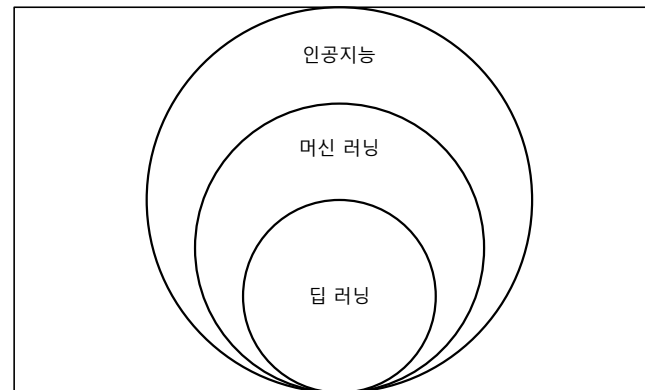
1.3. 딥러닝 의미

1.4 딥러닝 종류

1.5 딥러닝 역사

1.1 딥러닝, 머신러닝 관계

딥러닝에 대해 먼저 시작하기 앞서 머신러닝과 딥러닝의 관계를 알아보자. 머신러닝과 딥러닝은 인공지능 분야에 포함되는 기술이고 딥러닝은 머신러닝에 포함되는 기술이다. 인공지능과 머신러닝, 딥러닝의 관계를 표현하자면 <그림 1>과 같다.

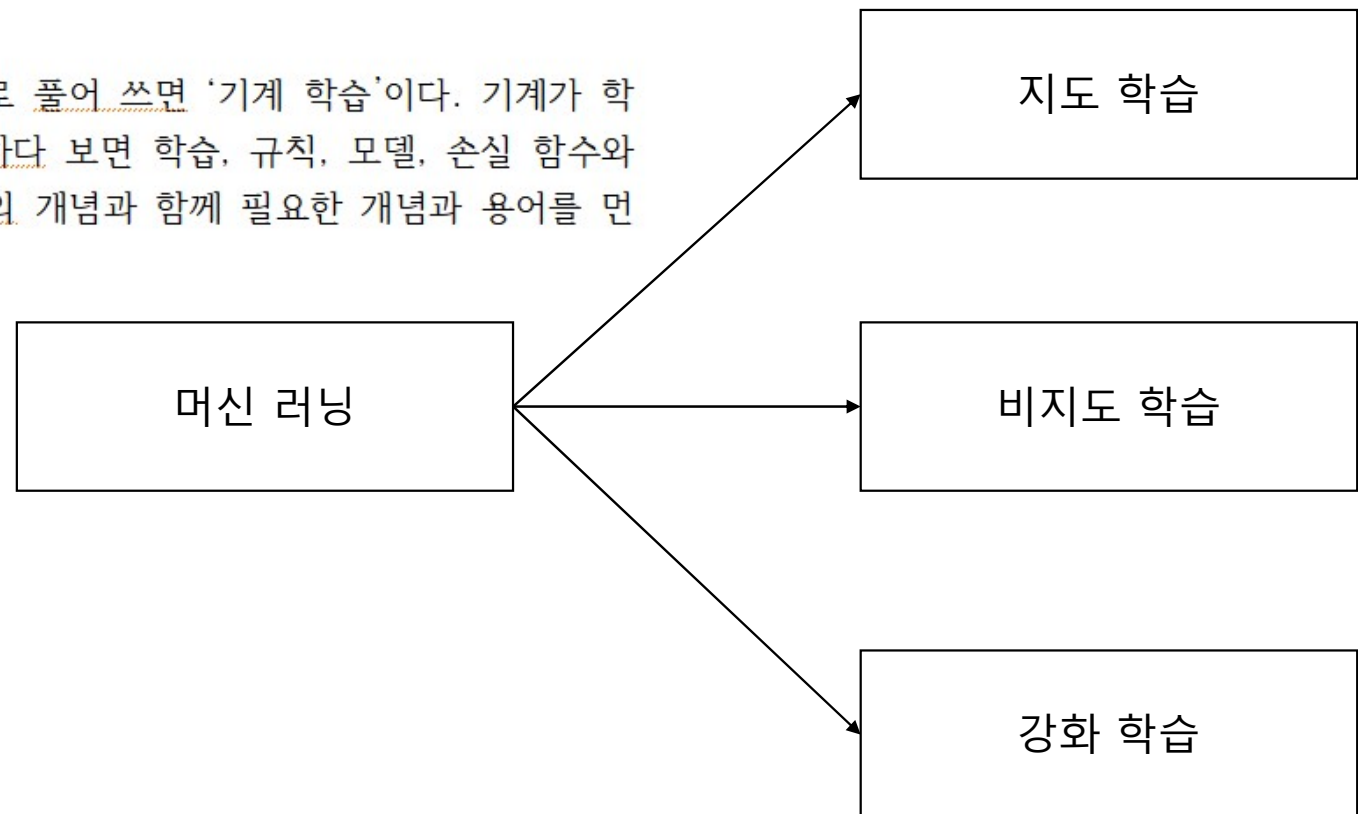


머신러닝은 알고리즘의 종류에 따라 더 세부적으로 나눌 수 있다. 그 중 인공신경망 알고리즘을 이용하여 만든 것이 딥러닝이다. 또한 인공신경망의 기본 구성 요소는 다른 머신러닝 알고리즘이 발전된 것이다. 딥러닝을 잘 이해하기 위해선 머신러닝을 먼저 알아야 한다. 따라서 머신러닝에 대해 먼저 설명을 한 후 딥러닝의 개념에 대하여 알아보자.

2. 2020년 10월 활동 보고서

1.2 머신러닝의 의미

머신러닝(Machine Learning)을 한글로 풀어 쓰면 '기계 학습'이다. 기계가 학습한다는 것은 어떤 뜻일까? 공부를 하다 보면 학습, 규칙, 모델, 손실 함수와 같은 용어들이 자주 나온다. 머신러닝의 개념과 함께 필요한 개념과 용어를 먼저 정리하고 넘어가려고 한다.



2. 2020년 10월 활동 보고서

지도 학습

다. 머신러닝으로 다루는 많은 작업들이 지도 학습에 속하는데 지도 학습이란 입력과 타겟으로 모델을 훈련시키는 것으로 입력에 대한 적절한 출력을 구하는 문제이다. 위 모델을 훈련 시키기 위해서 사용하는 데이터를 '훈련 데이터'라고 부른다. 그리고 훈련 데이터는 '입력'과 '타겟'으로 구성되어 있다. 입력은 모델이 풀어야 할 일종의 문제와 같은 것이며, 타겟은 모델이 맞춰야 할 정답과 같은 것이다. 문제에 대한 답을 주는 방법으로 모델을 훈련시키는 것이다. 학습을 통해 만들어진 프로그램을 모델이라고 하는데 모델은 새로운 입력에 대한 예측을 만든다. 따라서 지도 학습이란 기존의 데이터를 통해 모델을 학습 시키고 학습시킨 모델로 새로운 입력에 대한 예측을 할 수 있다. 따라서 지도 학습은 내일의 날씨를 예측하거나 스팸 이메일을 분류하는 등의 일을 해결할 때 많이 사용한다.

2. 2020년 10월 활동 보고서

비지도 학습

비지도 학습은 입력 정보의 특징을 찾는 문제이다. 따라서 비지도 학습은 타깃이 없는 데이터를 사용한다. 기업이 고객의 소비 성향에 따라 그룹을 지정하는 상황을 생각하면 비지도 학습의 개념을 조금 수월하게 이해할 수 있다. 그룹이 만들어지기 전까지는 어떤 그룹이 존재하는지, 몇 개의 그룹이 만들어질지 알 수 없다. 즉, 타깃이 없다. 타깃이 없으니 모델의 훈련 결과는 평가하기 어렵다는 특징이 있다. 비지도 학습의 대표적인 예로 군집(Clustering) 등이 있다.

강화 학습

강화 학습은 간단히 말하자면 장기나 체스와 같이 마지막 결과(전체적인 결과)가 가장 좋은 행동(예로 장기 말을 놓는 방법)을 찾는 문제이다. 강화 학습은 머신러닝 알고리즘으로 에이전트라는 것을 훈련을 시킨다. 훈련된 에이전트는 특정 환경에 최적화된 행동을 수행하고 수행에 대한 보상과 현재 상태를 받는다. 에이전트의 목표는 최대한 많은 보상을 받는 것이다. 따라서 에이전트는 주어진 환경에서 아주 많은 행동을 수행하여 학습이 된다. 대표적인 알고리즘으로는 Q-러닝(Q-Learning), SARSA 등이 있다. 강화 학습의 대표적인 예로는 딥마인드DeepMind의 알파고(AlphaGo) 등이 있다.

2. 2020년 10월 활동 보고서

자주 쓰이는 머신 러닝 용어

1 | 학습 : 머신러닝과 딥러닝에서 말하는 학습은 데이터의 규칙을 컴퓨터 스스로 찾아내는 것을 말한다.

2 | 규칙 : 전통적인 프로그램은 사람이 정한 규칙대로 동작하는 반면, 머신러닝은 사람이 만든 프로그램이지만 스스로 규칙을 찾아 수정한다.

3 | 모델 : 훈련 데이터로 학습된 머신러닝 알고리즘을 말한다.

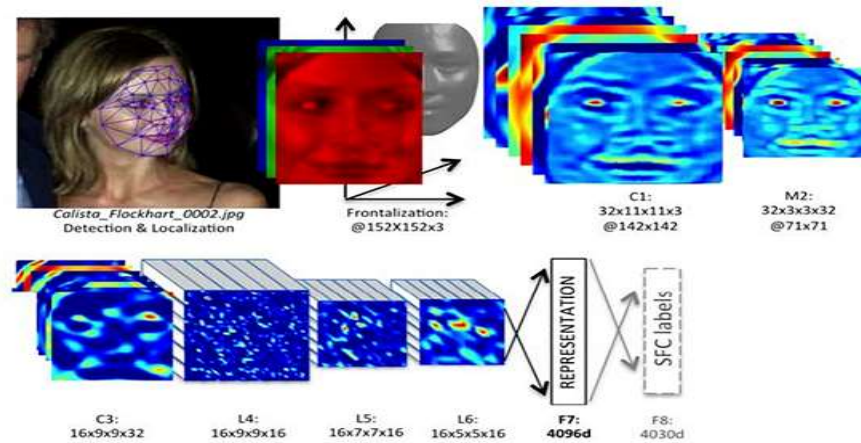
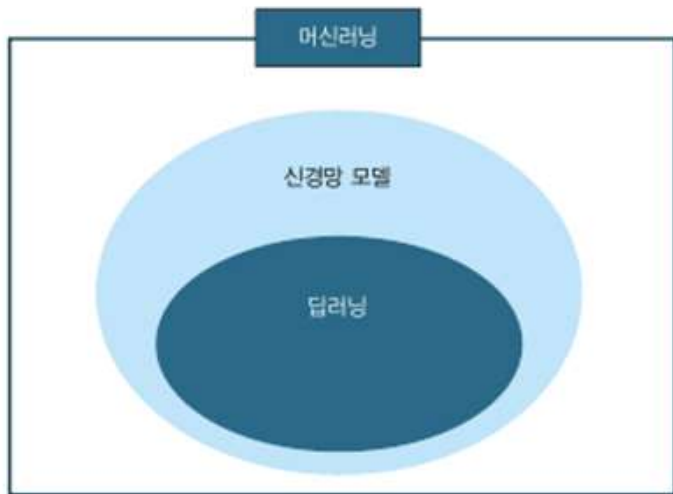
앞에서 만든 수학식이 바로 모델이다. 그리고 가중치와 절편을 합쳐 모델 파라미터(Model Parameter)이라고 부른다. 앞으로 이런 모델들을 파이썬으로 직접 구현해볼 것이다. 실제로는 모델을 클래스로 구현할 것이기 때문에 이를 통해 만든 객체를 모델이라고 생각하면 된다.

4 | 손실 함수 : 모델의 규칙을 수정하는 기준이 되는 함수를 말한다.

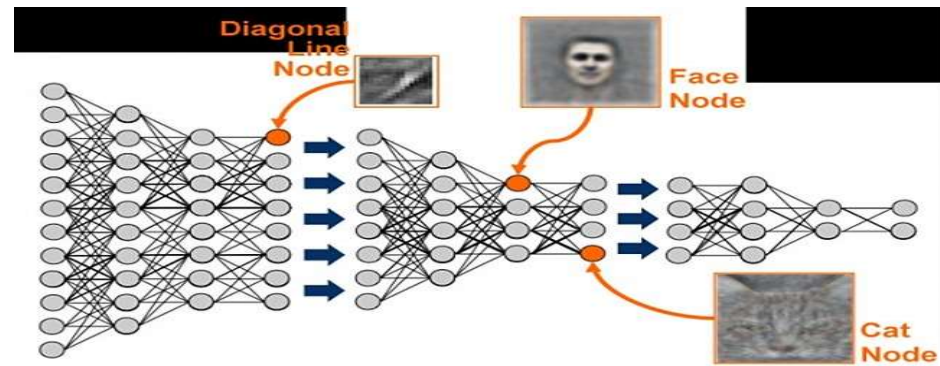
예를 들어 모델이 예측한 값과 타깃값 차이를 계산하는 함수를 손실 함수로 정의하면 차이가 작을수록 더 좋은 모델이라고 해석할 수 있다. 이때 최솟값을 효율적으로 찾는 방법을 최적화 알고리즘이라고 부른다.

2. 2020년 10월 활동 보고서

1.3 딥러닝 의미



<그림5> 페이스북의 '딥페이스' 동작 원리



<그림6> 유튜브 영상에서 고양이를 찾아내는 구글의 딥러닝 기술

2. 2020년 10월 활동 보고서

1.5 딥러닝 역사

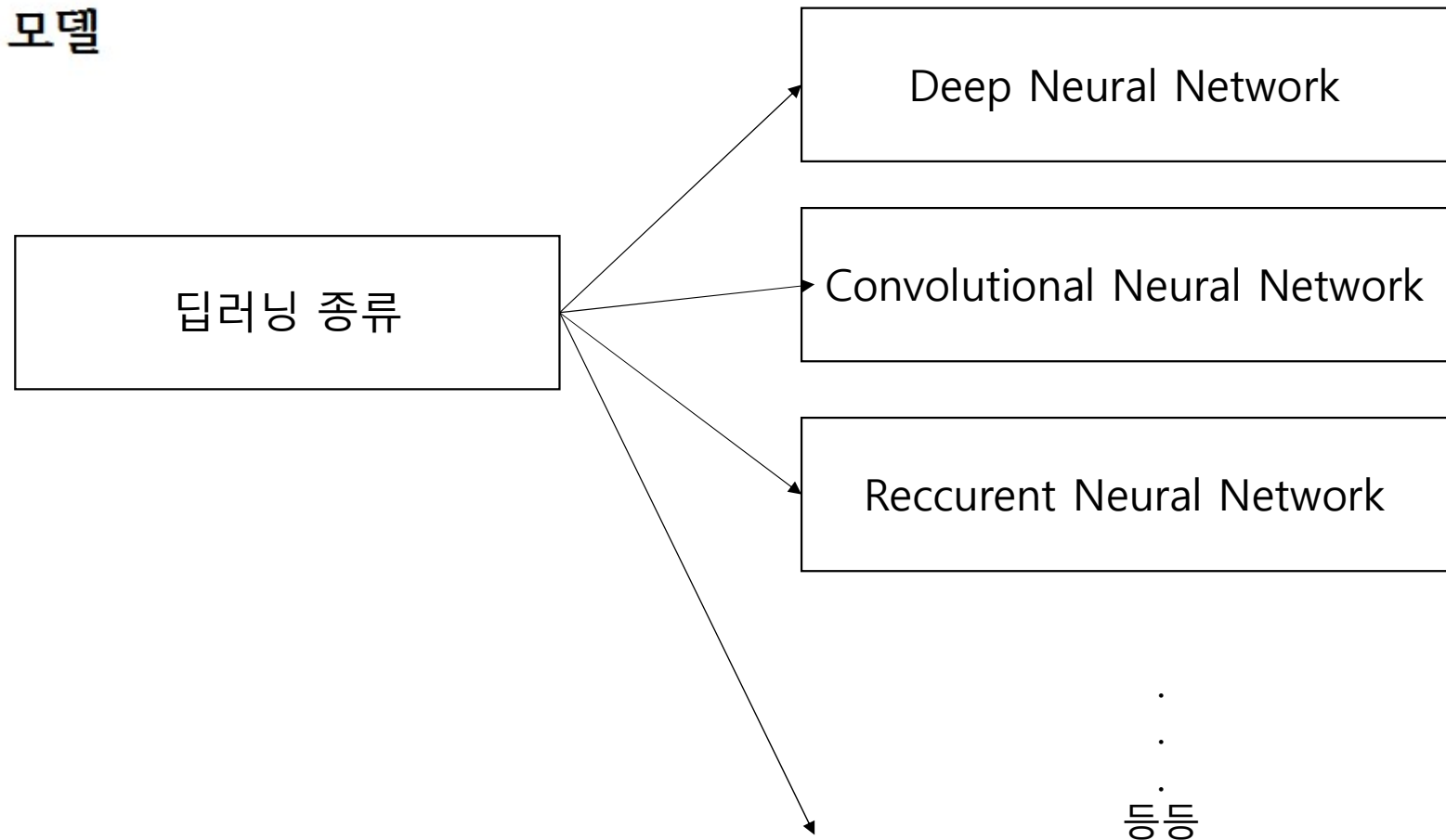
딥러닝 기법은 1980년 후쿠시마 쿠니히토(Kunihito Fukushima)가 소개한 신경망인 Neocognition에 처음 등장하였다. 1989년에는 얀 러쿤(Yann LeCun)과 동료들이 신경망에 표준 역전파(Backpropagation) 알고리즘을 적용한 연구를 수행하여 손으로 쓴 우편번호 인식에 성공하였으나, 신경망을 훈련시키는데 대략 3일이나 소요되었기 때문에 실용화 단계에는 이르지 못하였다. 신경망 연구에 있어서의 첫 번째 문제는 지나치게 긴 학습 시간이었고, 두 번째는 입력된 훈련 데이터에 과적합(Overfitting) 문제였다. 이러한 문제들로 인해 2000년까지는 신경망에 대한 연구보다 서포트 벡터 머신(support vector machine), k-최근접 이웃 알고리즘(K-Nearest Neighbor) 등의 단순한 머신러닝 알고리즘이 인기를 얻었다. 그러나 2006년 토론토대학교의 제프리 힌튼(Geoffrey Hinton)이 비지도 학습을 이용한 전처리 과정을 다층 신경망에 추가하는 방법을 통해 다층망을 쌓아도 정확성을 해치지 않는 방법을 개발하고, 2012년 개선된 기법을 객체 인식에 적용하여 오류율을 기존 방식 대비 10% 가량 떨어뜨리면서 딥러닝이 다시 주목을 받기 시작하였다. 2012년 스탠포드

대학교의 앤드류 응(Andrew Ng) 교수와 제프 딘(Jeff Dean)이 이끄는 구글 브레인 팀은 클라우드 환경을 기반으로 방대한 양의 유튜브 비디오를 자동으로 분석하여, 그 중 고양이의 이미지를 찾아내는 데 성공하였다. 여러 논문들이 나오며 지금은 다양한 분야로 확대되어서 연구되고 있다.

3. 2020년 11월 활동 보고서

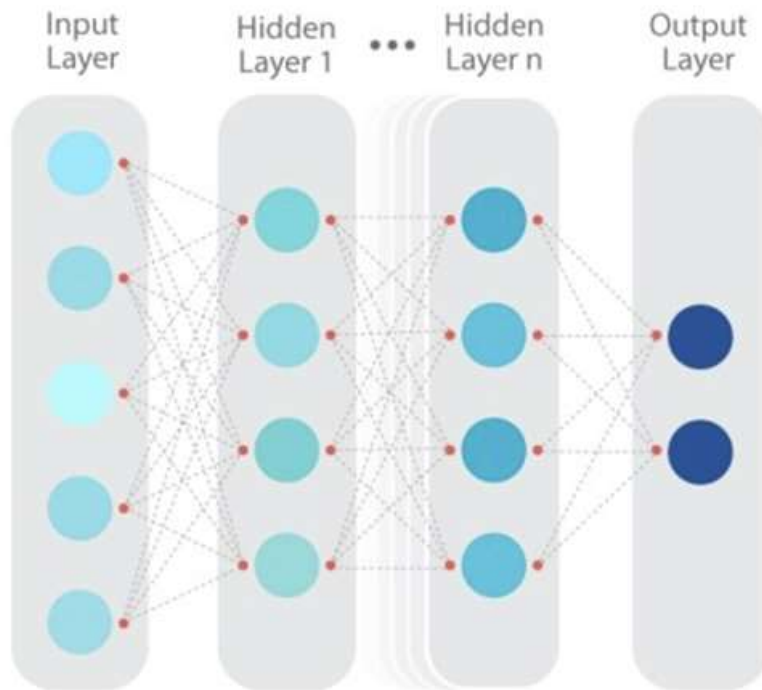
Chapter 2. 딥러닝 모델

- 2.1. 심층 신경망(DNN)
- 2.2. 합성곱 신경망(CNN)
- 2.3. 순환 신경망(RNN)
- 2.4. CTC Loss



3. 2020년 11월 활동 보고서

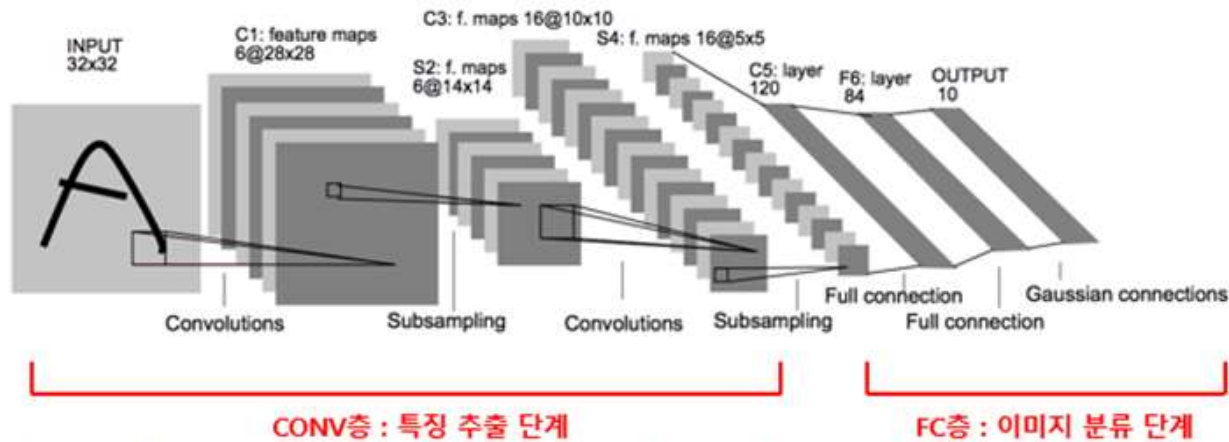
2.1. 심층 신경망(DNN)



입력층(input layer)과 출력층(output layer) 사이에 다중의 은닉층(hidden layer)을 포함하는 인공 신경망(Artificial Neural Network)이다. 심층 신경망(Deep Neural Network)은 다중의 은닉층을 포함하여 다양한 비선형적 관계를 학습할 수 있다. 그러나 학습을 위한 많은 연산량과 과하게 학습하여 실제 데이터에 대해 오차가 증가하는 과적합이나 기울기 값의 소실 문제 등의 문제가 발생할 수 있다. 2000년대 이후 다양한 기법이 적용되면서 딥 러닝의 핵심 모델로 활용되고 있다. 심층 신경망은 알고리즘에 따라 비지도 학습 방법(unsupervised learning)을 기반으로 하는 심층 신뢰 신경망(Deep Belief Network), 심층 오토인코더(Deep Autoencoder) 등이 있고, 이미지와 같은 2차원 데이터 처리를 위한 합성곱 신경망(Convolutional Neural Network), 시계열 데이터 처리를 위한 순환 신경망(RNN: Recurrent Neural Network) 등이 있다.

3. 2020년 11월 활동 보고서

2.2. 합성곱 신경망(CNN)



출처 : LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

CNN

Convolution layer

Pooling layer

Fully Connected layer

3. 2020년 11월 활동 보고서

합성곱 신경망(Convolutional Neural Network)은 입력된 이미지에서 다시 한번 특징을 추출하기 위해 kernel을 도입하는 기법이다. 심층 신경망(DNN)의 한 종류로, 크게 합성곱층(Convolution layer), 풀링층(Pooling layer) 그리고 완전연결층(Fully Connected layer)으로 구성된다. 합성곱층은 특징을 추출하는 단계이고 FC층은 이미지를 분류하는 단계이다. 먼저 합성곱층에 대해서 이야기를 해보자.

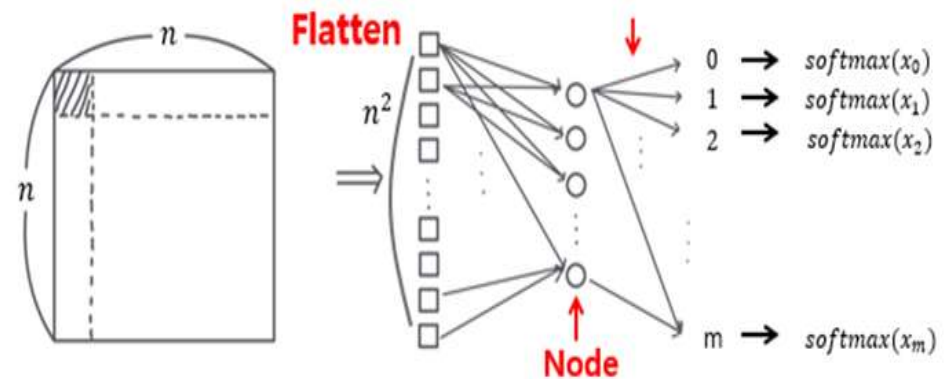
1. 합성곱층(convolutinal Layer)

이미지 특징을 추출하기 위해 우리는 합성곱층에서 필터를 씌운다. 필터란 이미지의 특징을 나타내기 위한 파라미터로 학습되는 대상이라고 생각을 하면 된다. 이미지를 넣었을 때 필터에 의해 합성곱으로 나온다.

2. 풀링층(Pooling layer)

풀링층은 데이터의 크기를 줄여 메모리를 작게 하고 과적합을 방지하는 역할을 하는 층이다. 풀링에도 다양한 종류가 있는데 가장 대표적인 풀링은 max pooling 으로 정해진 구간에서 최대값이 다른 특징들을 대표하는 방법이다.

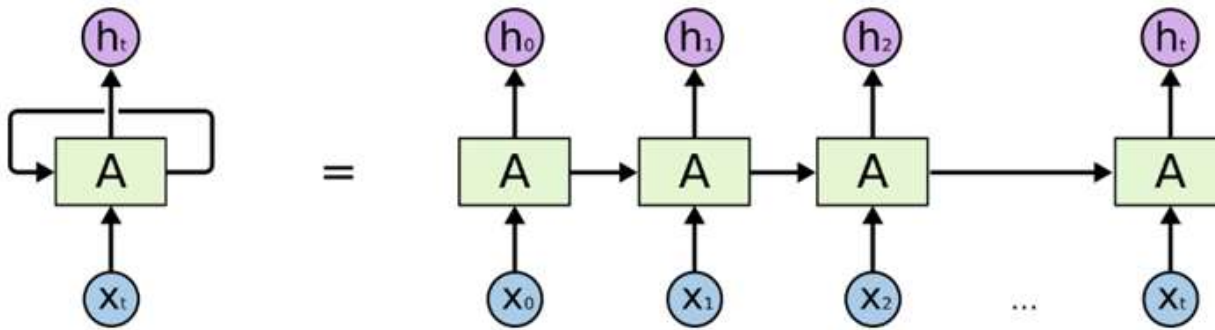
3. 완전연결계층(FC, Fully Connected layer)



완전연결계층은 앞의 과정을 거쳐 나온 데이터를 먼저 1차원 배열로 퍼준 다음에 층을 거쳐 결과를 내는 곳이다. 최종 분류를 위한 층이다.

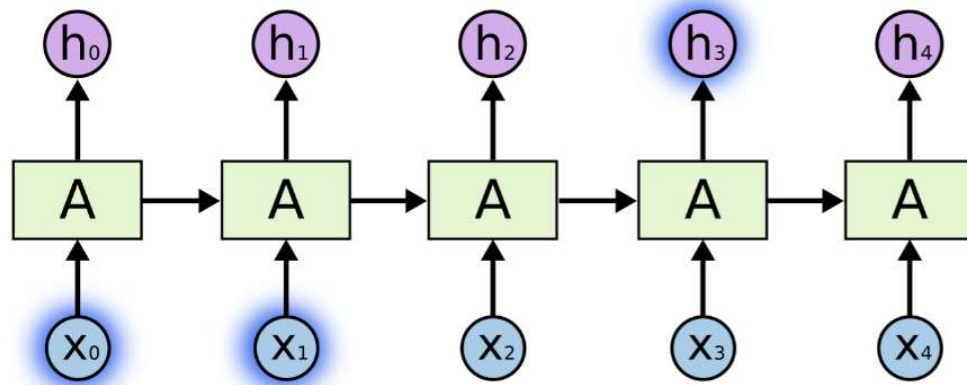
4. 2020년 12월 활동 보고서

2.3. 순환 신경망(RNN)



순환 신경망(Reccurent Neural Network)는 과거의 데이터가 미래에 영향을 줄 수 있는 구조를 가지고 있다. 사진에서 보이듯 A는 x_t 를 입력값으로 가지고 h_t 를 결과값으로 내놓는다. 이 회전은 정보가 다음 단계의 Neural Network로 이동하게 만들어준다. 이는 하나의 네트워크가 여러개 복사된 형태를 띄고 있다. 각각의 네트워크는 다음 단계로 정보를 넘겨준다, 따라서 연속이나 리스트에 관한 문제를 해결하기 위한 알고리즘으로 적절하다고 한다. 음성인식, 언어 모델링, 번역과 관련된 여러 분야에서 성공적으로 적용되어 성과를 내고 있다.

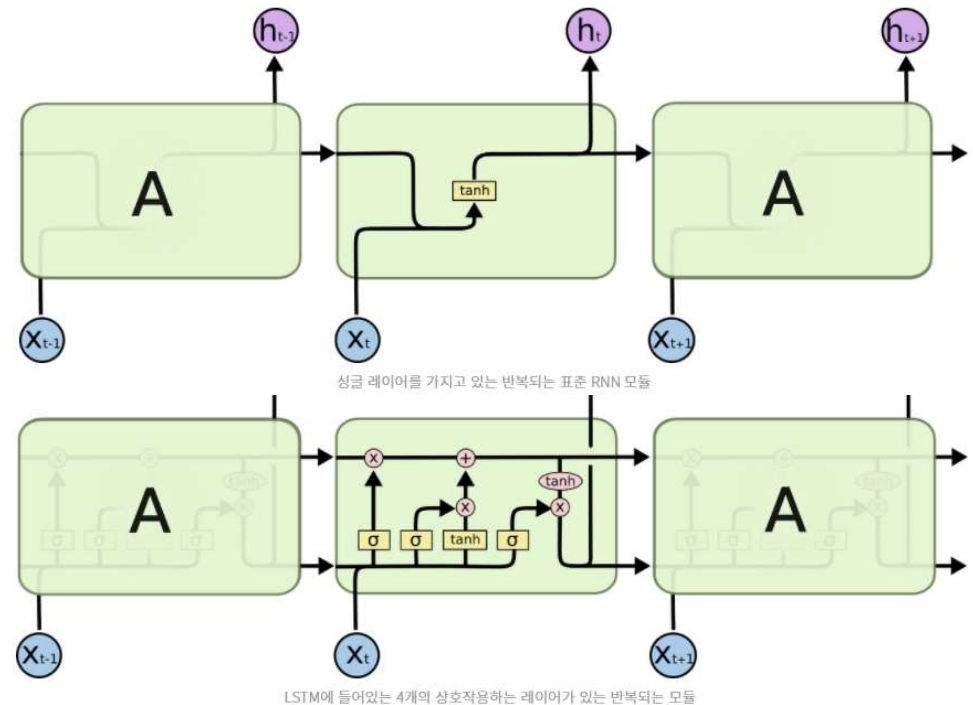
4. 2020년 12월 활동 보고서



순환 신경망의 장점 중 하나는 이전의 정보를 현재의 문제 해결에 활용할 수 있다는 점이다. 그런데 순환 신경망은 장기 의존성의 문제를 가지고 있다. 예를 들어 “the clouds are in the sky”라는 문장에서 “the clouds are the”라는 입력값을 받고 마지막 단어를 예측해야 한다면, 더 이상 문맥이 필요하지 않다. 다음에 입력될 단어는 “sky”가 될 확률이 높다. 이런 경우, 제공된 데이터와 배워야 할 정보의 입력 위치 차이가 크지 않다면 가능하다. 그렇지만 위의 사진처럼 배워야 할 정보의 입력 위치와 차이가 크다면 순환 신경망은 두 정보의 문맥을 연결하기 힘들어진다. 이러한 문제를 해결한 모델이 바로 LSTM이다.

4. 2020년 12월 활동 보고서

LSTM(Long Short Term Memory)은 순환 신경망의 종류이다. 장기 의존성 문제를 해결할 수 있다. LSTM은 Hochreiter & Schmidhuber (1997)이 제안하였고 많이 개선되고 대중화되면서 다양한 문제에 적용되기 시작했고, 지금은 많은 분야에서 사용하고 있다. 오랜 기간 동안 정보를 기억하는 일은 LSTM에 있어 특별한 작업 없이도 기본적으로 취하게 되는 기본 특성이다. LSTM과 RNN의 차이는 다음과 같다.



순환 신경망과 LSTM 모두 체인 구조를 가지고 있지만 반복되는 모듈은 다른 구조를 가지고 있다. 단일 Neural Network Layer를 가지는 대신에 4개의 상호작용 가능한 특별한 방식의 구조를 가지고 있다.

4. 2020년 12월 활동 보고서

2.4 CTC Loss

CTC(Connectionist Temporal Classification) loss는 조건부 확률의 음의 로그값이다. 조건부 확률을 나타내는데 CTC Alignment라는 특별한 방법을 사용하며 이는 이미지의 각 부분마다 어떤 글자가 있는지 정답을 알려주기 굉장히 힘들기 때문이다. CTC Alignment를 통해 이미지 X가 주어졌을 때 글자 Y가 나타날 조건부 확률을 계산하는 방법은 다음과 같다.

$$p(Y | X) = \sum_{L \in \text{map}(Y)} \prod_{t=1}^T p_t(l_t | X)$$

모든 Alignment에 대해서 확률을 더하는 방법은 동적 프로그램으로 구하게 된다.

5. 2021년 01월 활동 보고서

Chapter 3. Tesseract OCR 이해

3.1. OCR 개념

3.2. OCR 구조

3.3. Tesseract 개념

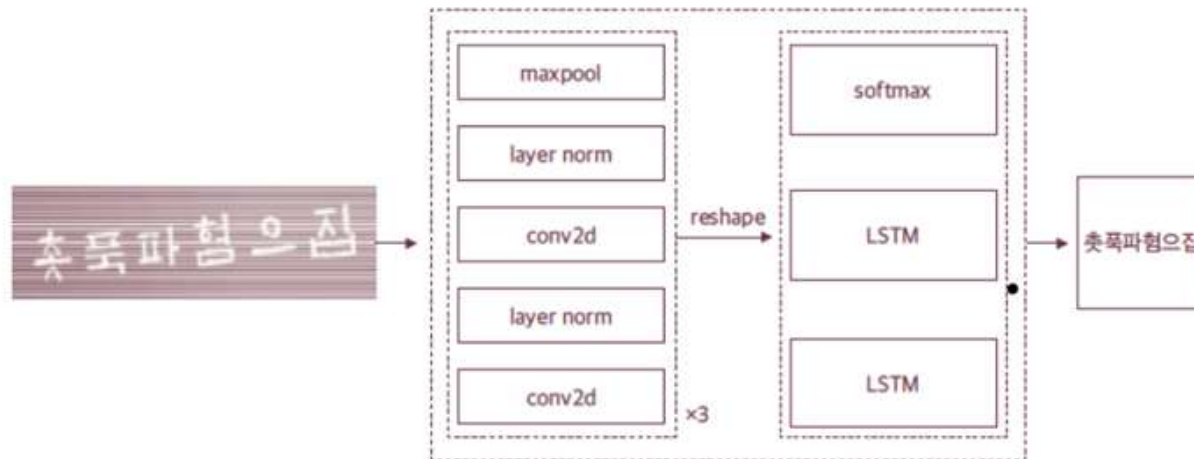
3.4. Tesseract 실습

3.1. OCR 개념

컴퓨터의 등장 이전에는 문서 기록을 위한 수단으로 주로 종이가 사용되었다. 이후 컴퓨터가 대중화되면서 종이 문서가 스캔을 통하여 전자 문서로 많이 변환되었다. 하지만 이렇게 변환된 전자 문서는 사람이 읽기만 가능하고 가공하지 못한다. 가공을 위해서는 변환된 전자 문서를 사용자가 편집 가능한 텍스트 문서로 바꾸어야 한다. 이를 사람의 수작업으로 수행한다면 노동의 낭비와 비효율, 그리고 처리 가능한 양의 한계에 부딪히게 된다. 이러한 문제는 이미지 파일에 존재하는 글자를 편집 가능한 텍스트 형태로 바꾸어 주는 OCR(Optical Character Reader) 기술을 통해 많은 부분 해결되었다. OCR이란 광학 문자 판독기로 문서에 새겨진 문자를 빛을 이용해 판독하는 장치를 말한다. 즉, 문자, 숫자 또는 다른 기호의 형태가 갖는 정보로부터 디지털 컴퓨터에 알맞은 부호화된 전기신호로 변환하는 장치이다. 주로 특정 형태의 타이핑된 문자를 판독하는 것이 많다. 이 OCR 기술은 카드 인식, 우편물 분류 등 다양한 분야에서 활용되고 있다.

5. 2021년 01월 활동 보고서

3.2 OCR 구조



<그림> 글자 인식 모델의 구조

OCR 모델은 합성곱 신경망, 순환 신경망, CTC 알고리즘으로 구성되어 있다.

5. 2021년 01월 활동 보고서

3.3 Tesseract 개념

Tesseract는 Apache 2.0 라이선스에 따라 사용할 수 있는 OCR 엔진이다. 현재 공식 배포는 4.1.1 버전이다. 하지만 실습할 버전은 alpha 5.0.0이다. Tesseract에 대한 최신 코드를 원하는 사람들은 Github에서 다운 받아서 사용할 수 있다. Tesseract는 명령 줄을 통해 직접 사용하거나 API를 사용하여 이미지에서 인쇄된 텍스트를 추출할 수 있다. 또한 다양한 프로그래밍 언어를 지원한다.

다양한 유형의 모델에 대한 자세한 정보는 홈페이지를 참고하면 된다.

구글 테세라트 공식 홈페이지 <https://tesseract-ocr.github.io/tessdoc/>

1

Tesseract-OCR 설치

2

윈도우에서 실행

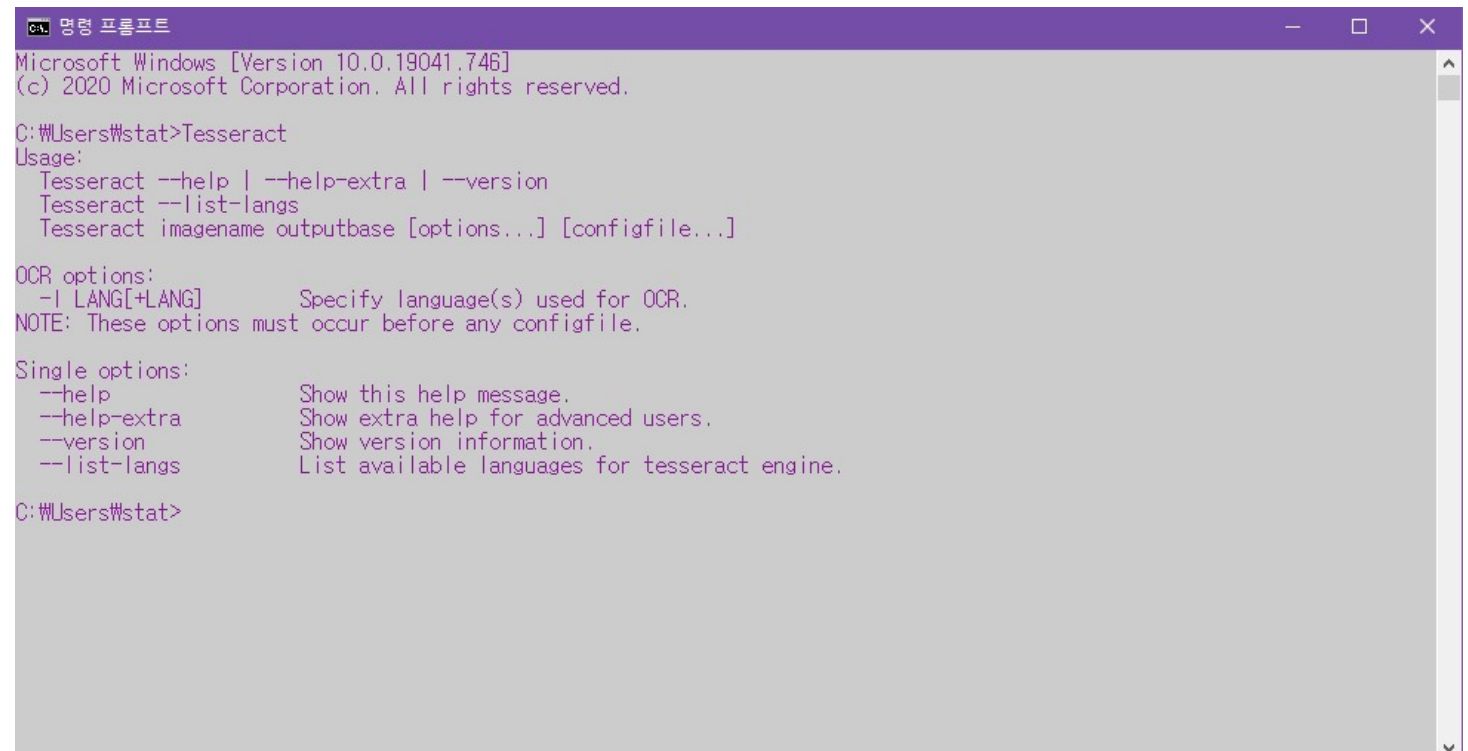
3

Python에서 사용

5. 2021년 01월 활동 보고서

1

Tesseract-OCR 설치



```
C:\> 명령 프롬프트
Microsoft Windows [Version 10.0.19041.746]
(c) 2020 Microsoft Corporation. All rights reserved.

C:\Users\stat>Tesseract
Usage:
  Tesseract --help | --help-extra | --version
  Tesseract --list-langs
  Tesseract imagename outputbase [options...] [configfile...]

OCR options:
  -l LANG[+LANG]      Specify language(s) used for OCR.
NOTE: These options must occur before any configfile.

Single options:
  --help              Show this help message.
  --help-extra        Show extra help for advanced users.
  --version           Show version information.
  --list-langs        List available languages for tesseract engine.

C:\Users\stat>
```

5. 2021년 01월 활동 보고서

2

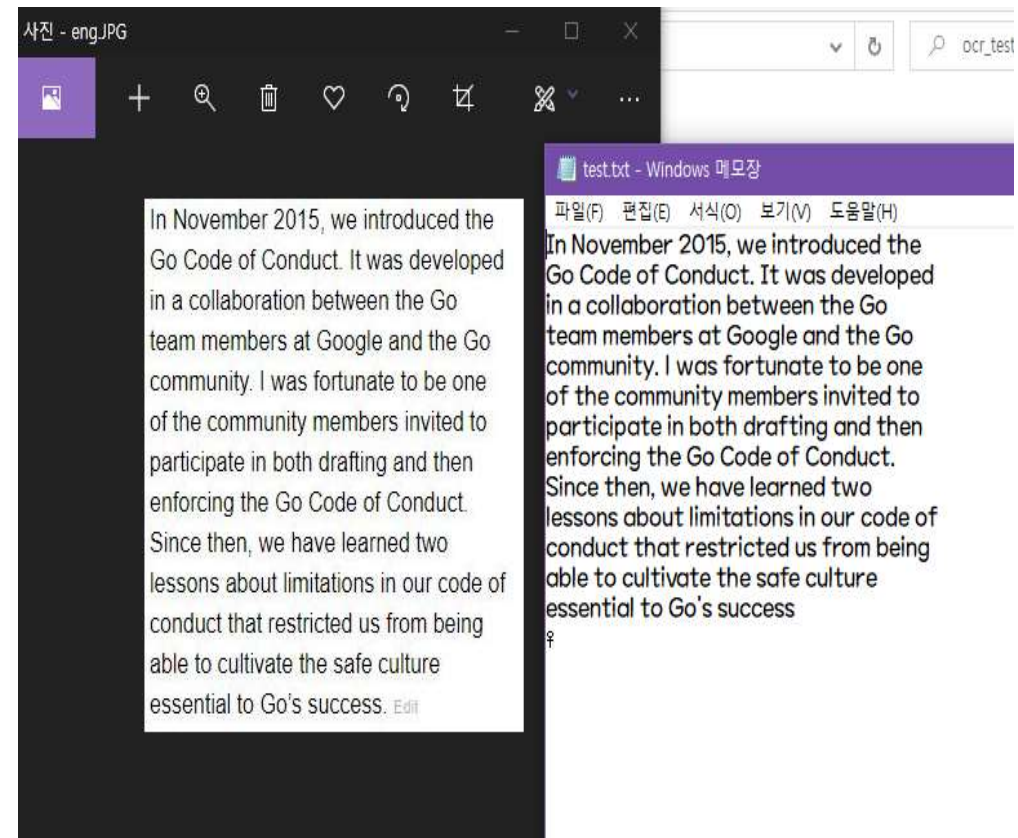
윈도우에서 실행

```
tesseract imagename outputbase [options...] [configfile...]
```

```
선택 명령 프롬프트
--help-extra Show extra help for advanced users.
--version Show version information.
--list-langs List available languages for tesseract engine.

C:\Users\#stat>tesseract "C:\Users\#stat\Desktop\#저지해#R_D 인턴십#ocr_test#eng.JPG" "C:\Users\#stat\Desktop\#저지해#R_D 인턴십#ocr_test#test"
Tesseract Open Source OCR Engine v5.0.0-alpha.20201127 with Leptonica

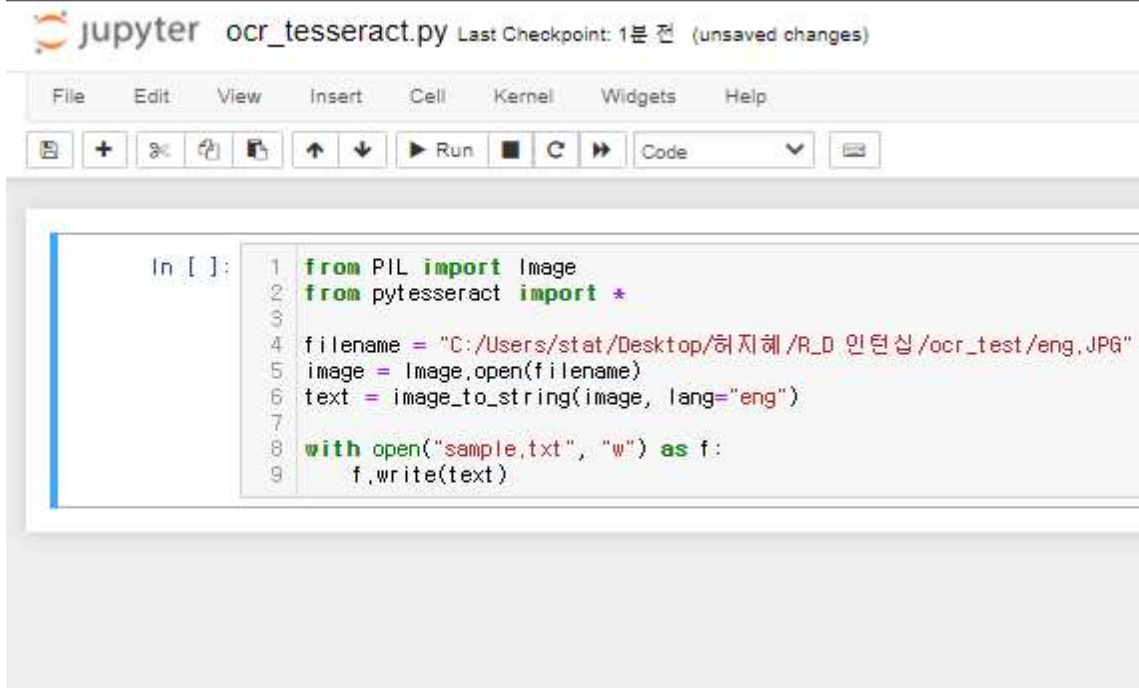
C:\Users\#stat>
```



5. 2021년 01월 활동 보고서

3

Python에서 사용



```
jupyter ocr_tesseract.py Last Checkpoint: 1분 전 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

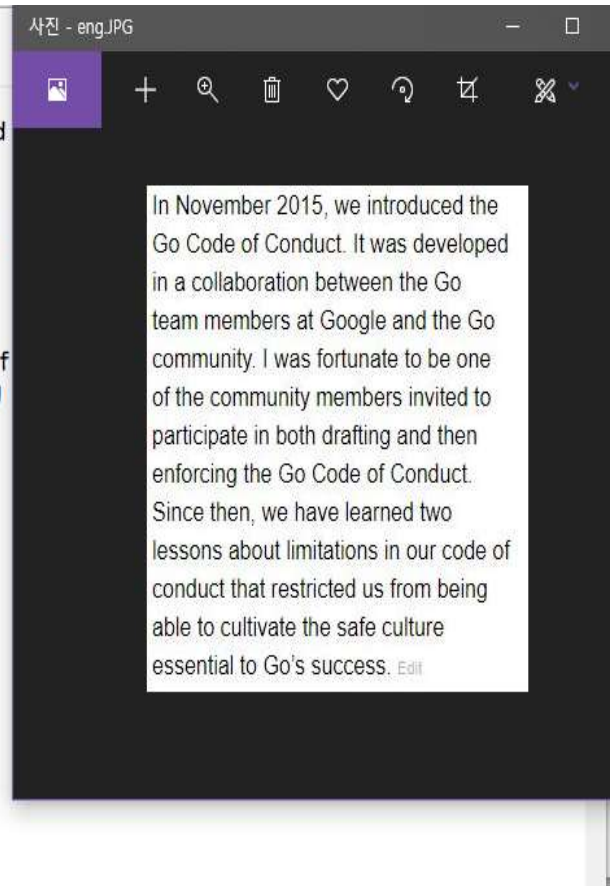
In [ ]: 1 from PIL import Image
        2 from pytesseract import *
        3
        4 filename = "C:/Users/stat/Desktop/허지혜/R_D 민원실/ocr_test/eng.JPG"
        5 image = Image.open(filename)
        6 text = image_to_string(image, lang="eng")
        7
        8 with open("sample.txt", "w") as f:
        9     f.write(text)
```



sample.txt - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

In November 2015, we introduced the Go Code of Conduct. It was developed in a collaboration between the Go team members at Google and the Go community. I was fortunate to be one of the community members invited to participate in both drafting and then enforcing the Go Code of Conduct. Since then, we have learned two lessons about limitations in our code of conduct that restricted us from being able to cultivate the safe culture essential to Go's success.



6. 결론 & 느낀 점



하고 싶은 공부를 계획대로 할 수 있어서 좋았다. 데이터셋도 만들었지만 적용하는 과정은 완성을 못했다. 하지만 세미나를 통해서 부족한 부분을 채우고 Tesseract를 실행했기 때문에 목표는 이뤘다. R&D 인턴십 기회가 또 주어진다면 더 잘할 수 있을 것 같다.

출처

- 모두의 딥러닝 - (주)도서출판길벗 조태호 저
- 파이썬으로 배우는 머신러닝의 교과서 - (주)한빛미디어, 아코 마토코 저/ 박광수(아크몬드)역
- Do it! 정직하게 코딩하며 배우는 딥러닝 입문 - (주)이지스퍼블리싱 박혜선 저
- [네이버 지식백과]
- 딥러닝 - 학습을 통한 생각하는 컴퓨터 (용어로 보는 IT, 오원석),
- 딥 러닝 [Deep Learning] (두산백과),
- 콘볼루션 신경망 [Convolutional Neural Network, -神經網] (IT용어사전, 한국정보통신기술협회),
- 심층 신경망 [Deep Neural Network, 深層神經網] (IT용어사전, 한국정보통신기술협회),
- ocr 네이버 지식백과 시사상식사전
- 딥러닝을 이용한 한글 OCR 정확도 향상에 대한 연구 논문
<https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=NPAP12688214&dbt=NPAP>
- LSTM, RNN 개념 및 사진 <https://brunch.co.kr/@chris-song/9>
- OCR 글자인식모델구조 <https://brunch.co.kr/@kakao-it/318>
- 구글 Tesseract 개념 <https://tesseract-ocr.github.io/tessdoc/>
- Tesseract 실습 <https://joyhong.tistory.com/79>

끝