머신러닝 기반의 대규모 이미지 파일에서 개인 정보 분류 시스템

김기태*, 윤상혁O, 서보인*, 이세훈*
OO인하공업전문대학 컴퓨터시스템과,
*인하공업전문대학 컴퓨터시스템과
e-mail: kkt@inhatc.ac.kr, yshg97@gmail.com, sbi1024@naver.com, seihoon@inhatc.ac.kr

Machine Learning based Personal Information Classification System in Large Image Files

Ki-Tae Kim*, Sang-Hyeok YunO, Bo-in Seo*, Sei-hoon Lee*
ODept. of Computer System, INHA TECHNICAL COLLEGE,
*Dept. of Computer System, INHA TECHNICAL COLLEGE

● 요 약 ●

본 논문에서는 현재 이슈가 되고 있는 개인 정보 보안에 대해서 Keras 라이브러리를 사용하여 개인 정보 관련 데이터를 학습한 후, 한글 인식률 증가된 Tesseract-OCR 활용하여 사람들이 가지고 있는 데이터의 개인 정보 유무를 판단하여 분류한다.

키워드: 머신러닝(Machine Learning), 개인정보분류(Personal Information Classifying), OCR(optical character recognition)

I. Introduction

직장이나 업무 현장 같은 곳은 문서, 팩스 등 개인 정보를 가지고 있는 데이터들이 많이 존재한다. 보직 이동을 하거나 퇴직을 할 때는 개인 정보가 들어있는 파일들을 폐기해야 하는데, 그럴 때마다 개인 정보를 가지고 있는 지를 사람이 일일이 검사를 하고 있다.

이에 본 논문에서는 이러한 과정을 머신러닝과 OCR을 이용하여 쉽게 자동화 하는 방법을 제시한다. 그리고 이미지의 개인 정보 유무를 판단하기 위해 글자로 추출할 때 Tesseract-OCR을 사용하였는데, 이때 지원하는 한글 언어 데이터의 한글 인식률이 떨어지는 현상이 존재한다. 이에 본 논문에서는 Tesseract-OCR의 한글 인식률 증가 방안도 제시한다.

II. System Architecture

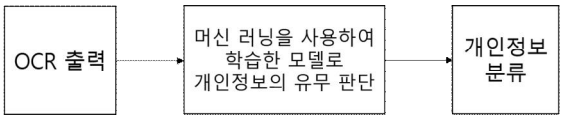

Fig. 1. System Architecture

Fig 1은 개인 정보 분류 시스템 구성도이다. 본 시스템은 Keras로 개인 정보 형식에 들어가는 특징 텍스트 데이터를 학습 시킨다. 그리고 Tesseract로 이미지 파일을 OCR로 출력하고 이미지의 글자를 추출한 것을 활용하여 개인 정보 데이터를 학습 시킨 모델로 OCR 출력한 결과물이 개인 정보를 가지고 있는지, 없는지를 판단하고 이에 따라 개인 정보를 가지고 있다면 그 파일을 지정해 놓은 디렉터리로 분류시킨다.

III. Experiments

1. 한글 데이터 폰트별 Fine Tune학습

Tesseract-OCR의 한글 데이터 인식률을 증가시키기 위해 ubuntu 18.04LTS 환경에서 Tesseract의 한글 학습 데이터의 Fine Tune 학습을 진행한다.[1] 폰트는 굴림체, 바탕체, THEPure 체(한국 그림 필기체), 네이버 나눔 고딕 4종의 폰트데이터와 국립국어원에서 한국어 학습용 어휘 목록 데이터(5965개 단어)를 추가하여 학습한다.

2. 한글 데이터 학습 과정

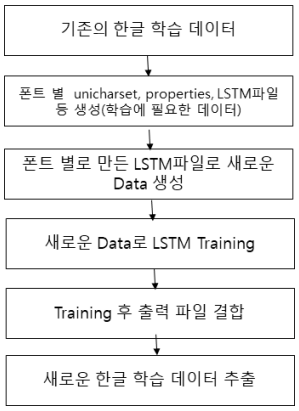

Fig. 2. Learning Process

Fig 2은 기존의 한글 학습 데이터를 폰트 별 학습 시키는 과정을 나타낸다. 기존의 한글 학습 데이터를 복사한다. 그리고 추가한 폰트를 가지고 학습에 필요한 데이터(unicharset, unichar properties, lstm 파일)를 만든다. 기존 한글 학습이 되어 있는 데이터를 가져와서 새로운 데이터로 변환한다. 그리고 원하는 폰트별로 만든 새로운 데이터들로 lstm training을 한다. 학습을 한 후 출력 파일을 결합한다. 최후에는 새로운 한글 학습 데이터를 추출할 수 있다.

```
At iteration 0, stage 0, Eval Char error rate=0.76648772
, word error rate=1.7739105
```

Fig. 3. When learned first fonts

```
At iteration 0, stage 0, Eval Char error rate=0.59267039
, word error rate=1.3288057
```

Fig. 4. When learned fourth fonts

Fig 3, Fig 4에서는 Tesseract에서 제공하는 모델 확인용 텍스트 파일을 OCR 출력하여 비교한 후 오탈자율을 확인한 것을 나타낸다. 학습을 할수록 Char error rate와 word error rate의 값이 1.77에서 1.32로 작아지는 것을 확인할 수 있다.

3. Keras를 이용한 개인 정보 텍스트 학습

Fig 5는 Keras를 사용한 개인 정보 텍스트 데이터의 전처리와 모델 학습의 순서도이다.

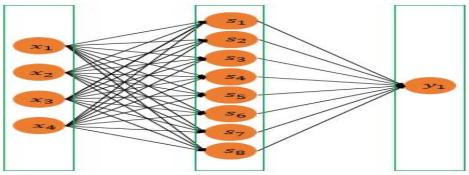

Fig. 5. Text Classification using Keras

이미지를 OCR추출한 데이터의 개인 정보 유무를 판단하기 위해 Keras 토크나이저를 통해 각각 단어에 토큰화와 정수 인코딩 과정을 수행하여 빈도수가 높은 순서대로 단어에 인덱스를 부여한다. 그리고 평균 글자 수를 기준으로 패딩 작업 진행하여 모든 데이터의 길이를 맞춰 주고 모델을 학습시킨다.

4. 학습한 모델로 이미지의 개인 정보 유무 판별하여 분류

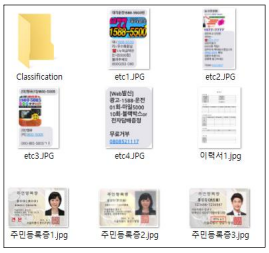

Fig 6. Files with personal information and files without personal information

Fig 7. categorized only images with personal information

Fig 6는 개인 정보가 있는 이미지와 개인 정보가 없는 이미지를 보여주는 화면이다. Fig 7은 개인 정보가 있는 이미지(주민등록증, 이력서)들이 분류되어 다른 경로에 저장된 화면이다. 이미지를 OCR 추출하고 그 결과 텍스트를 학습된 모델로 평가하였을 때 사용자가 지정한 수치(80%) 이상이면 개인 정보가 있는 파일로 분류되는 것을 확인할 수 있다.

IV. Conclusions

본 논문은 Tesseract OCR의 한글을 출력했을 때 인식률의 향상시키는 방법과 이미지를 OCR 출력하여 그 이미지에 대한 개인 정보 유무 판단과 분류에 대해 제안한다. 한글 학습 데이터를 더 학습시키고 개인 정보의 특징을 세분화하여 학습시켜 개인 정보의 데이터의 분류를 더욱 개선이 가능하다.

REFERENCES

[1] https://tesseract-ocr.github.io/tessdoc/TrainingTesseract-4.00.html