Portfolio

데이터분석 희망자 이해강

ABOUT ME

동아대학교 경영정보학과 이해강 입니다.
4학년에 데이터 분석에 흥미를 가지고
데이터 분석에 관한 공부를 하였습니다.
현재 python를 통한 CNN, LSTM,
KoGPT2, 딥러닝 알고리즘을 실습 통해
응용력을 키웠습니다.

신세계와 동아대에서 주관하는
SW 인재양성 프로그램에서
3번의 프로젝트를 통해 팀원 간의
소통법과 프로젝트에 필요한
알고리즘을 활용 할 수 있게 되었습니다.

My Github

https://github.com/leehg9805/leehg9805.git

1. CNN 피부암 예측 모델

CNN 피부암 예측모델

소속 동아대 경영정보학과

작 성 1718285 이해강

자 1951959 김정수

주제 CNN을 활용한 피부암 예측모델

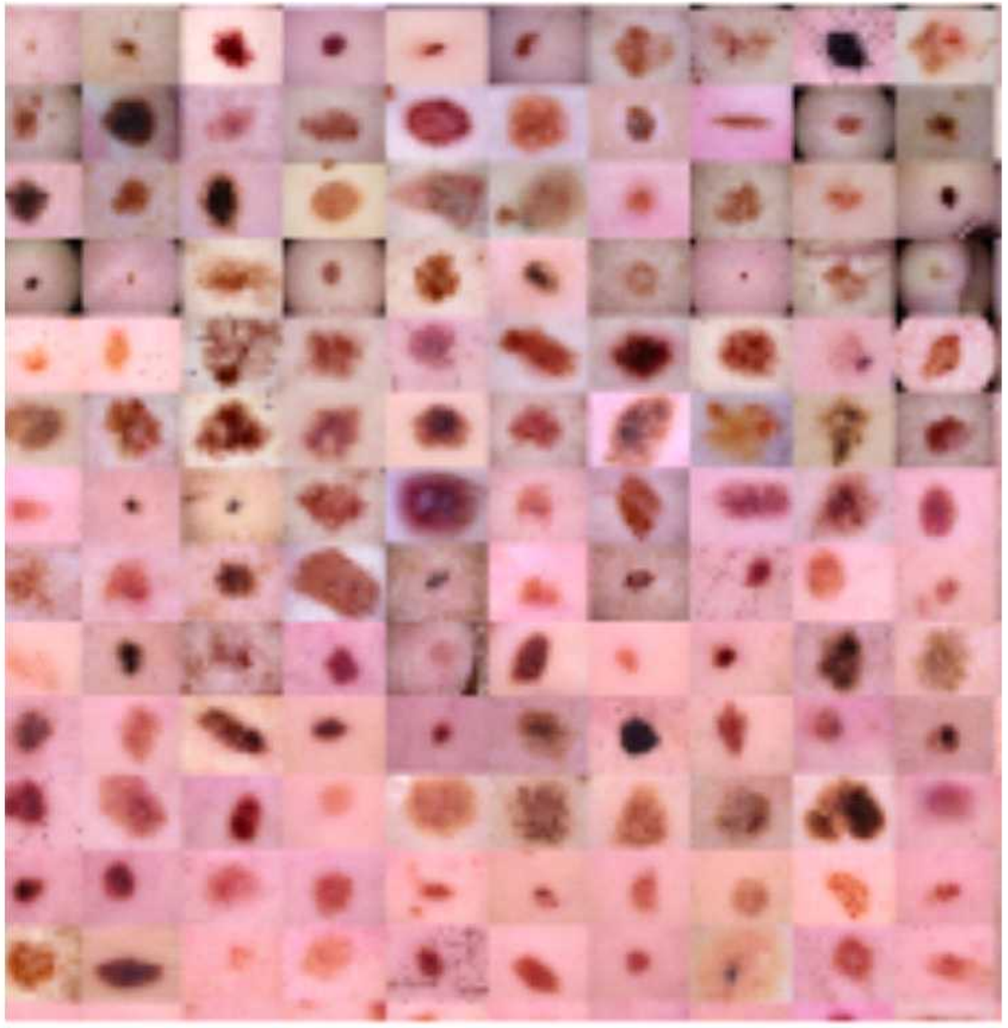
1) 개요

현재 우리나라 2020년 피부암 발생환자 수는 27,211명으로 2016년 19,236에 비해 5년 동안 41.5% 증가, 전체 암 발생의 0.5%를 차지하였고 인구 10만 명당 조발생률은 2.6명으로 보고되었다.(보건복지부 중암암등록본부 2012년 12월 발표 자료)

피부암 환자 대부분이 고령층인 점, 거동이 불편한 점, 초기에 증상이 거의 없기 때문에 예방하기 어려운 점을 고려해 피부암 발생위험에 대한 선별적 검사를 시행하여 고위험군을 선별하고 본인의 병명을 조기에 알 수 있게 알려줌으로써 피부암의 초기 예방, 피부암으로 인한 사망을 줄이는 것이 중요합니다.

그렇기에 이미지 학습에 적합한 CNN을 활용한 인공지능 모델을 만들고자 해당 프로젝트를 시작 하였습니다.

1. CNN 피부암 예측 모델



2) 맡은 업무

딥러닝을 하기 위해서 필요한 이미지를 kaggle(Skin Cancer MNIST:HAM10000)에서 가져온 후 CNN 모델에서 학습을 할 수 있도록 전처리를 하였습니다.(경로설정, 필요 없는 열 삭제)

좋은 모델을 찾기 위해서 총 네가지의 방식으로 층을 다르게 쌓아서 학습을 한 모델의 혼동행렬의 비교를 통해 좋은 모델을 선정하였습니다.

결과물을 토대로 학습과정과 결론을 보고서에 작성을 하였습니다.

1. CNN 피부암 예측 모델

3) 결과

	precision	recall	f1-score	support
0	0.79	0.43	0.56	44
1	0.69	0.68	0.69	60
2	0.61	0.57	0.59	109
3	0.56	0.50	0.53	10
4	0.50	0.40	0.45	99
5	0.85	0.93	0.89	663
6	1.00	0.29	0.45	17
accuracy			0.79	1002
macro avg	0.72	0.54	0.59	1002
weighted avg	0.78	0.79	0.78	1002

	precision	recall	f1-score	support
0	0.75	0.20	0.32	44
1	0.84	0.52	0.64	60
2	0.60	0.52	0.56	109
3	0.60	0.30	0.40	10
4	0.40	0.62	0.49	99
5	0.86	0.91	0.89	663
6	0.91	0.59	0.71	17
accuracy			0.77	1002
macro avg	0.71	0.52	0.57	1002
weighted avg	0.79	0.77	0.77	1002

	precision	recall	f1-score	support
0	0.77	0.39	0.52	44
1	0.68	0.63	0.66	60
2	0.65	0.54	0.59	109
3	0.38	0.50	0.43	10
4	0.39	0.53	0.45	99
5	0.88	0.89	0.88	663
6	0.91	0.59	0.71	17
accuracy			0.77	1002
macro avg	0.67	0.58	0.61	1002
weighted avg	0.78	0.77	0.77	1002

	precision	recall	f1-score	support
0	0.58	0.34	0.43	44
1	0.75	0.40	0.52	60
2	0.50	0.62	0.55	109
3	0.36	0.50	0.42	10
4	0.53	0.41	0.46	99
5	0.86	0.92	0.89	663
6	0.90	0.53	0.67	17
accuracy			0.77	1002
macro avg	0.64	0.53	0.56	1002
weighted avg	0.77	0.77	0.76	1002

본 프로젝트를 간략히 요약하면 합성곱 신경망(콘볼루션 신경망, Convolutional neural network)을 이용하여 피부암 예측모델을 개발하는 것이다. 모델의 층은 4가지의 방식으로 쌓았으며, 개발한 피부암 예측모델을 이용한 성능 확인을 위해 각 모델의 정확도를 분석해 보았을 때 모델1 0.9205, 모델2 0.9052, 모델3 0.9100 모델4 0.7768로 모델1의 정확도가 가장 높음을 알 수 있었습니다.

모델1은 신경망을 128, 64개로 주고 합성곱층을 3개로 쌓았다. 또한 전체적인 학습결과를 비교를 하였을 때 정확도가 몇 번 지점에서 가장 높은지에 대한 정보만 얻을 수 있을 뿐 어느 모델이 더 뛰어난지에 대해서는 정확히 알아낼 수 없었습니다.

단, 모델1과 모델3은 전이학습 전에는 과적합이 일어나는 모습을 보여주었지만 전이학습을 한 후에는 과적합이 줄어든 모습을 볼 수 있었습니다.

성공적인 결과를 얻음으로써 본 논문의 연구가 이론적인 내용만을 다루는 것이 아니라 실제 적용이 가능함을증명하였다 볼 수 있었습니다. 다만 이미지를 끌어온 KAGGLE는 자료의 불균형을 이루고 있는 상태라 위 결과가 완벽하다 단정지을 순 없다 이후 연구에서는 개발한 프로그램을 더 다양한 방법과 자료를 토대로 보완해 나가는 추가 연구가 필요하다는 결론을 얻게 되었습니다.

2. 내 손안의 닥터봇

1) 개요

첫째, 고령의 사람들이 자신의 병을 진료를 할 수 있는 진료과를 인지를 하지 못해 진료과를 돌아다니는 불편함을 줄이기 위해서

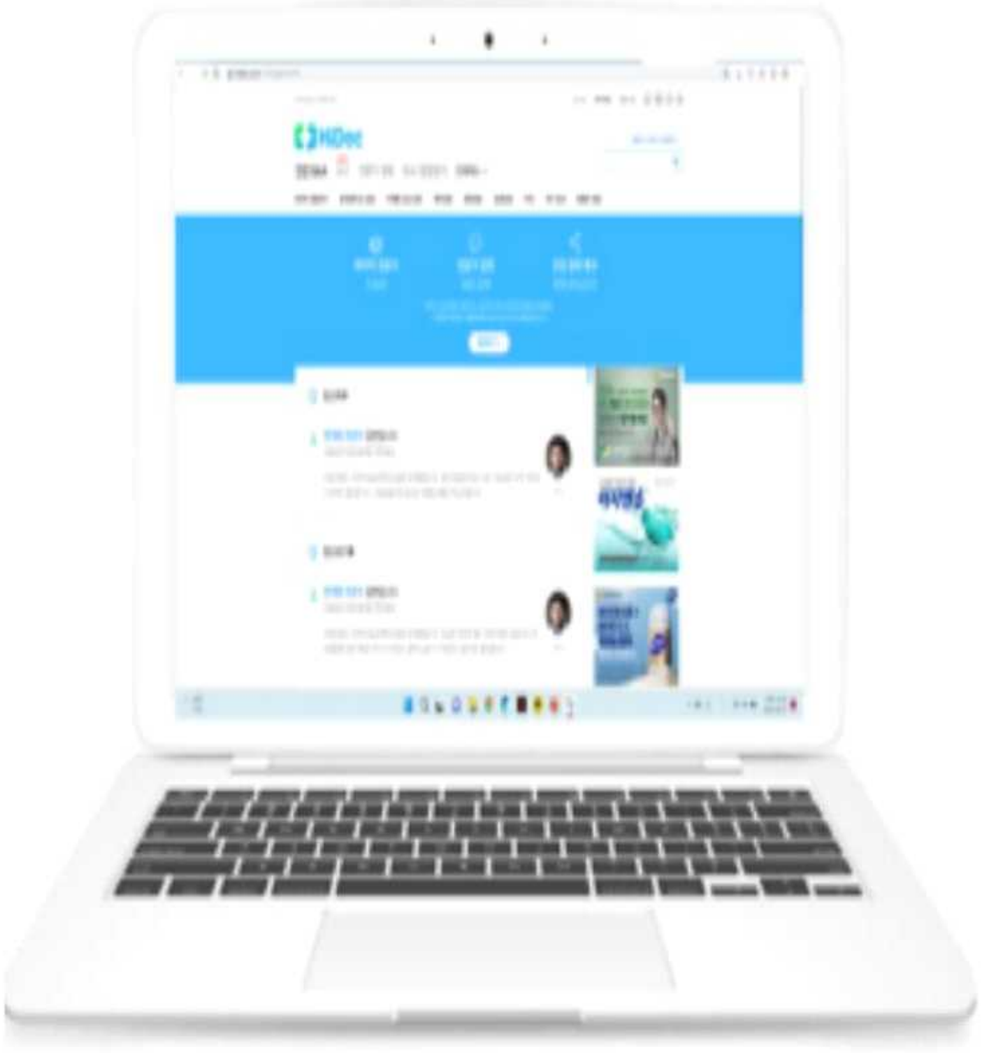
둘째, 자신의 증상을 인터넷에 찾아 보는 것으로 병명을 판단을 하려는 사람들이 있다고 한다. 하지만 인터넷은 정보가 많은 만큼 불확신한 정보가 많기에 증명된 정보를제공하기 위해서

셋째, 초기에 증상을 잡아야지 치료를 할 수 있는 병을잡아내기 위해서

넷째, 고령자의 의료형평성 미확보에 따른 지속적인 의료취약계층이 생겨나고 있어서

위의 내용을 종합을 하였을 때 증상을 입력을 하였을 때병명을 알려주는 챗봇을 만드는 것이 적합하다고 판단을하였습니다. 그리고 정확한 정보를 전달을 하기 위해서전문가의 소견을 볼 수 있는 사이트인 하이닥과 네이버 지식인을 통해서 정보를 수집하였습니다.

2. 내 손안의 닥터봇



2) 맡은 업무

정보를 수집을 하기 위해서 전문가의 소견이 들어간 진단을 볼 수 있는 하이닥 사이트에서 webdrive를 이용하여 환자의 증상과 병명을 크롤링 하였습니다.

모은 자료는 학습을 할 수 있도록 확실하게 증상만 입력이 된 자료만 남기도록 전처리를 하였습니다.

Jupyter를 통해서 그래픽카드를 연동을 시키고 LSTM를 이용하여 챗봇에 활용 할 수 있는 모델을 제작 TTS를 통해서 음성인식도 가능한 모델을 제작하였습니다.

2. 내 손안의 닥터봇

3) 결과

정확도로 평가를 했을 때 RNN(0.63), LSTM(0.74), KoBERT(0.76), KoGPT2(0.82)로 KoGPT가 높은 정확도를 나타냈다. 하지만 대장염, 독감, 협심증 증상을 평가를 했을 때 KoBERT가 좋은 결과를 가져왔습니다.

이러한 결과는 KoBERT와 KoGPT2에 저장 되어있는 단어 사전의 차이라고 보고 있습니다.

총 2,4000개의 데이터를 수집을 하였지만 모델을 완성을 시키기에는 부족하다고 판단을 하였다. 그렇기에 추후에는 병원과 협업을 토대로 완성을 시키는 방향을 잡았습니다.

3. 피부질환 디텍션을 통한 스킨케어

1) 개요

국내 스킨케어 시장이 규모가 증가하고 스킨케어 관련 키워드 검색량이 증가함에 따라 소비자들의 관심이 커지는 것을 알 수 있었습니다. 하지만 스킨케어에 관심을 가지고 있다고 하더라도 잘 못 된 정보를 통한 민간요법으로 오히려 피부를 악화가 되는 경우가 있다고 합니다.

그렇기에 해당 프로젝트를 통해 피부질환을 파악을 하고 피부질환에 필요한 성분과 치료법을 피부질환 환자에게 알려줄 수 있는 인공지능을 만들었습니다.

3. 피부질환 디텍션을 통한 스킨케어

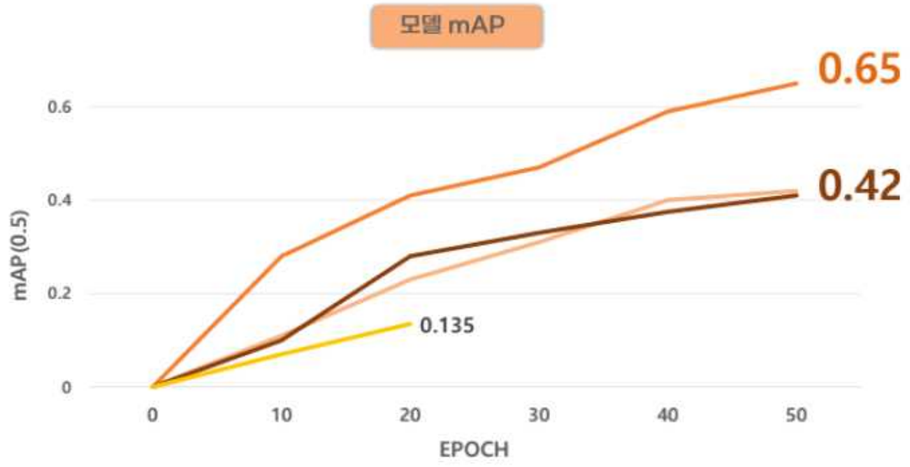
2) 맡은 업무

학습을 시키기 위해 필요한 이미지를 모으기 위해 구글, 네이버를 통해 한국어, 영어, 일본어, 중국어로 구글링을 하여 총 2만 1천개의 이미지를 수집을 하였습니다.

yolov7과 Faster-R-cnn를 이용하기 위해서 이미지를 라벨링을 하는 것으로 전처리를 진행했습니다.

RPN를 통해 RoI 계산이 가능하게 하여 기존 R-CNN 모델보다 높은 정확도를 보유하고 있으며 정적 detection에 좋은 모델인 Faster-R-cnn를 이용하여 인공지능 모델을 만들었습니다.

3. 피부질환 디테션을 통한 스킨케어

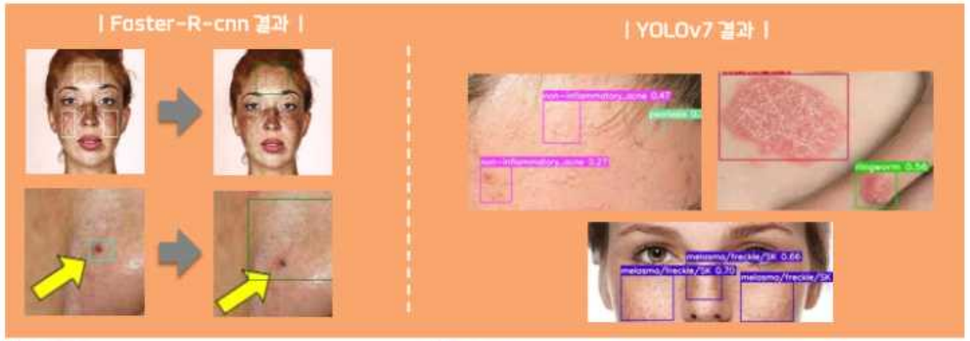


3) 결과

예상으로는 Faster-R-cnn이 정확도가 높다고 생각을 하였지만 정확도 확인 결과 Faster-R-cnn보다 yolov7의 정확도가 높은 것을 확이 하였습니다.

정확도로는 확실한 결과를 파악을 할 수 없어 만들어진 모델을 이용하여 테스트를 진행 한 결과 Faster-R-cnn 보다 yolov7이 더 많이 맞춘 것을 확인 하였습니다. 이러한 결과를 통해 R-cnn 기반 모델이 비슷한 객체를 구별을 못 한다는 것을 확인 하였습니다.

yolov7는 Faster-R-cnn보다 좋은 결과를 얻었지만 모든 질환을 맞추지는 못 하였기에 좀 더 많은 이미지들을 모아 학습을 시키는 것으로 모델을 향상시키기로 하였습니다.



한계점
- cnn 기반으로 비슷한 물체를 잘 잡아내지 못함

개선방향
- 주제와 맞는 코드 개발

한계점
- 데이터 부족과 다양하지 못 하여 제대로 된 학습이 되지 않음

개선방향
- 많은 데이터를 이용하여 모델을 학습 하여야 함

감사합니다.