4차 6기 세미프로젝트

IDC FINDER

2020년 3월 23일

딥러닝 기반 AI 엔지니어링 (A)

SOMAC

팀장 이찬호

팀원 이정철

정소현

정용주

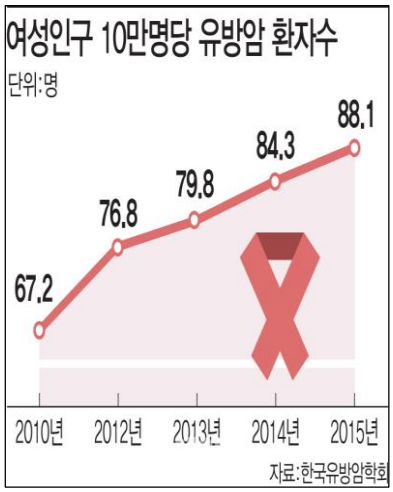
황지민

목 차

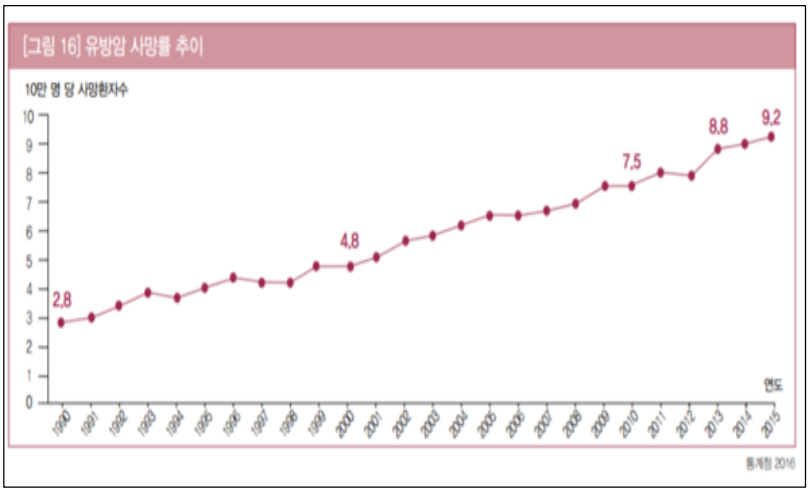
1. 프로젝트 개요

1.1 프로젝트 기획 배경 및 목표

유방암은 전 세계 여성 암의 25.2%를 차지할 만큼 발생빈도가 높은 질병이다 (출처 : 2018 유방암 백서). 미국 등에서 자주 발생하는 선진국형 질병이나, 최근 서구화된 식습관과 고령화로 인해 국내 유방암 환자도 크게 증가하고 있다. 한국유방암학회의 자료에 따르면 여성인구 10 만명당 전체 유방암 환자 수는 2000 년 26.3 명에서 2015 년 88.1 명으로 지속적으로 증가했다 <그림 1> . 특히 우리나라는 폐경 전 유방암 환자 비중이 46.5% (2015 년 기준)에 달할 정도로 젊은 유방암 환자 비중이 높은 편에 속한다. 또한 통계청에 따르면 유방암으로 인한 사망률은 여성 인구 10 만 명 당 4.8 명이었던 2010 년에 비해, 2015 년엔 약 2 배 늘어난 9.2 명을 기록하였다 <그림 2>. 사망에까지 이르지 않더라도, 전조증상이 거의 없다는 점, 폐 전이와 뼈 전이가 흔히 나타난다는 점, 치료 과정에서 유방 절제가 필요하다는 점, 재발률이 높다는 점 등의 이유 때문에 유방암은 많은 여성들이 두려워하는 질병이다.



<그림 1>



<그림 2>

유방암 중에서도 침윤성 유관암(Invasive Ductal Carcinoma)은 전체 유방암의 약 75-85%를 차지하는 가장 대표적인 유방암이다. 따라서 교육과정 중에 배웠던 이미지 분류(ImageNet)를 통해 침윤성 유관암 여부를 판별함으로써 정확한 유방암 조기진단에 도움이 되고자 하여 프로젝트 주제를 IDC(Invasive Ductal Carcinoma) FINDER 로 정하게 되었다.

1.2 구성원 및 역할

이름	전공	역할	구현 부분
이찬호	회계	팀장	프로젝트 총괄 및 일정관리 팀원 업무 분담 데이터 분석
이정철	경영	팀원	Backend 개발 (Flask)
정소현	통계	팀원	데이터 분석
정용주	전자	팀원	Frontend 개발
황지민	기계	팀원	데이터 분석 Backend 개발 (Flask)

1.3 프로젝트 추진 일정

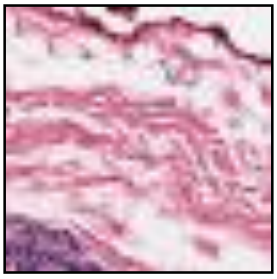
구분	기간	활동	비고
사전 기획	2020-03-17	프로젝트 기획 및 팀 구성	
	2020-03-17	PJT 주제 선정, 팀(PM/팀원) 구성	5 인/팀
PJT 수행 / 완료	2020-03-17	데이터 전처리	
	2020-03-18 ~ 2020-03-19	케라스 모델 구현	
	2020-03-18 ~ 2020-03-19	Frontend 구현	
	2020-03-20	Frontend 보완	
	2020-03-20	Flask api 서버 구현	
	2020-03-23	팀 별 최종 발표 (구축 완료 보고)	

2. 프로젝트 개발 결과

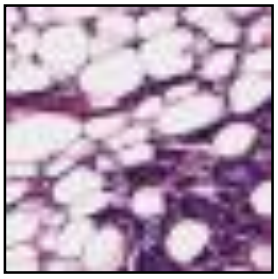
2.1 데이터 수집

캐글에서 2 년 전 진행되었던 'Breast Histopathology Images' 대회의 데이터 셋을 프로젝트 데이터로 이용하였다. 이 데이터 셋은 40 배 확대해서 스캔한 유방암 검체의 슬라이드 이미지 279 개에서 추출된 50X50 픽셀 크기의 27 만 7524 개의 이미지(19 만 8738 개의 음성 이미지, 7 만 8786 개의 양성 이미지)로 구성되어 있다.

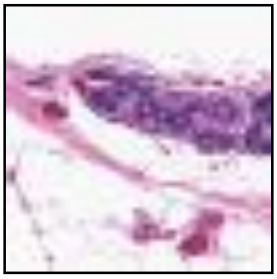
다음은 임의로 추출한 환자의 음성 이미지와 양성 이미지이다.



<그림 3> 8863 환자의 음성 이미지



<그림 4> 8863 환자의 양성 이미지



<그림 5> 10253 환자의 음성 이미지

<그림 6> 10253 환자의 양성 이미지

<그림 7> 16896 환자의 음성 이미지

<그림 8> 16896 환자의 양성 이미지

2.2 데이터 분석

Data Sampling

데이터 셋은 음성 이미지가 19만 8738개, 양성 이미지가 7만 8786개로 그 비율이 약 7:3 정도이다. 이러한 클래스 불균형에서 오는 over-fitting을 방지하기 위해 음성 이미지 중 랜덤으로 7만 8786개만을 샘플링 하여 양성 이미지 개수와의 비율을 맞추어 주었다. 이렇게 조정한 15만 7572개의 이미지를 Sklearn 라이브러리의 train_test_split() 함수를 이용해 랜덤하게 Train Data(80%)와 Validation Data(10%), Test Data(10%)로 분류하였다.

Data Preprocessing

모델이 학습 데이터에만 맞춰지는 Over-Fitting을 방지하기 위해서, 모델이 이미지에서 최대한 많은 정보를 학습하게 하는 Image Argumentation 과정이 필요하다. Keras 라이브러리에서 제공하는 ImageDataGenerator 클래스를 이용해 이러한 전처리를 수행할 수 있다. 이를 통해 이미지를 회전, 평행이동, 반전, 확대, 축소 등의 과정을 거쳐 train data를 늘려주었다.

Data Modeling

자원의 한계와 예측 시간의 절약을 위해, 작은 크기이고 낮은 계산 복잡도를 갖지만 정확도의 손실도 최소화하는 효율적인 모델인 MobileNetV2를 이용했다. Sequential 객체를 생성해 MobileNetV2을 대입하고, 그 결과를 Flatten()을 이용해 1차원 배열로 변환하였다. 그 후에 Dense를 이용한 선형 회귀와, Over-Fitting을 방지하기 위한 Dropout 작업을 수행해 모델을 만들고, Train Data를 학습시켰다.

Prediction

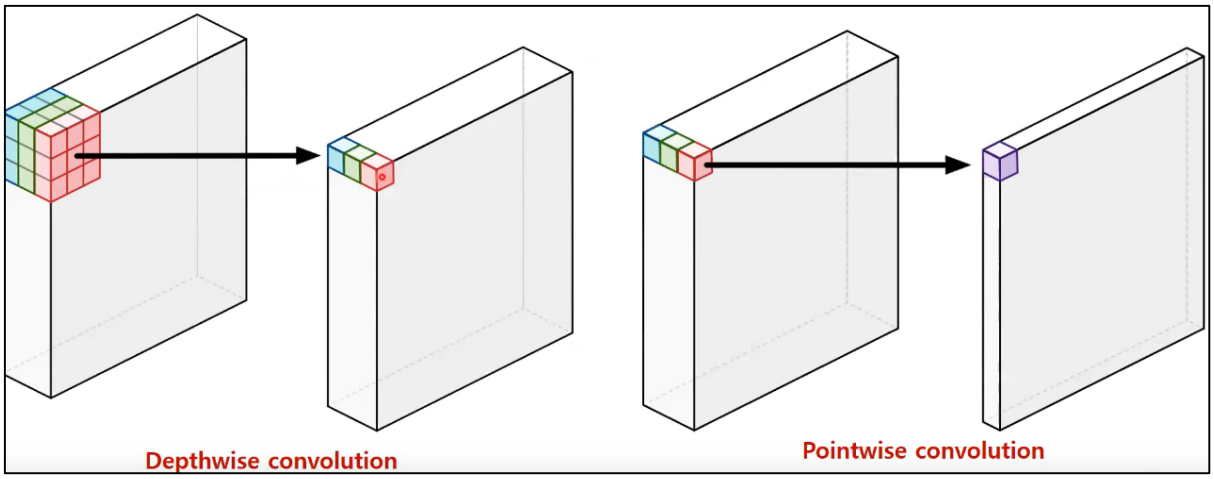
Train Data 학습을 완료한 모델을 Test Data로 정확도를 계산하였다. Predict() 함수를 이용해 확률값을 리턴 받고, 그 값이 0.5 미만이면 음성 (0으로 Labeling), 0.5 이상이면 양성 (1으로 Labeling)으로 변환하여 Test Data의 Label 값과 비교하였다. 비교 결과 True이면 1, False이면 0으로 변환한 값을 평균을 내어 확률을 계산하였다.

2.3 데이터 분석 결과

핵심 모델

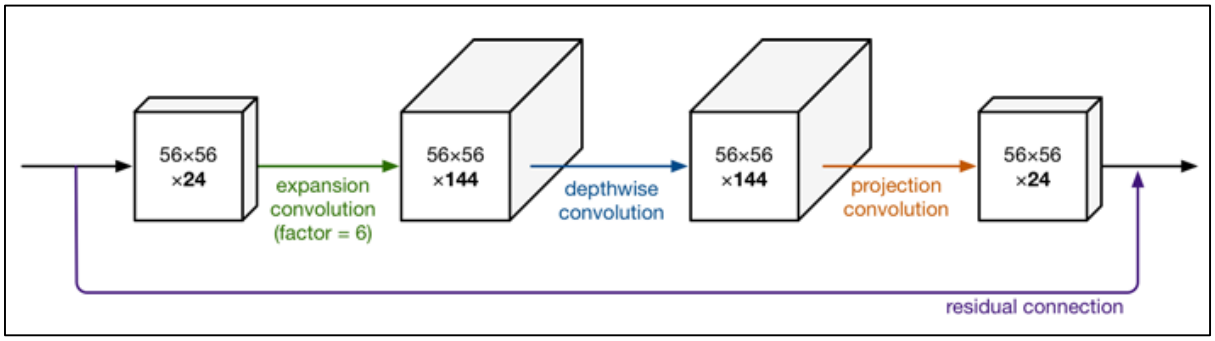
Train Data 를 학습시키는 데 사용했던 주된 모델이 MobileNetV2 이다. MobileNetV2 는 크게 Depthwise Separable Convolution 과 Linear Bottleneck 을 이용해서 만들어진 네트워크이다.

Depthwise Separable Convolution 방식은 전체의 channel 방향과 spatial 한 방향 모두를 한꺼번에 고려한 기존의 convolution 방식과 달리, 이 둘을 완전히 분리하겠다는 아이디어이다 <그림 9>. 이렇게 두 방법으로 분리를 하여도 channel 과 spatial 한 방향 모두 보기에, 기존의 convolution 과 유사하게 작동되나 parameter 의 수와 연산량이 기존의 방법보다 훨씬 적다. 수치로 따지면 대략 8~9 배의 수준으로 계산량이 줄어든다.



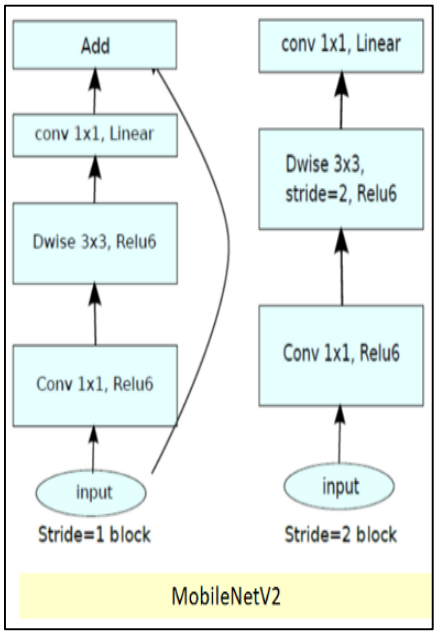
<그림 9>

Linear Bottleneck 은 ReLU 함수를 사용하면 비선형에 의한 정보 손실이 일어나지만, 차원 수가 충분히 큰 공간에서 ReLU 함수를 사용하는 경우에는 이러한 정보 손실율이 낮아지게 된다는 사실에서 시작된다 <그림 10>. 따라서 MobileNetV2 모델은 데이터 손실을 최소화하기 위해서 저차원의 데이터를 expansion 층을 통해 확장시키고 depthwise convolution 연산을 거쳐 마지막에 projection 층을 통해 많은 데이터를 보존할 수 있도록 설계되어 있다.



<그림 10>

이러한 두 가지 특성 때문에 MobileNetV2 는 낮은 계산 복잡도를 가지는 작은 모델이지만 정확도의 손실은 최소화하는 효율적인 모델이라고 할 수 있다. 이러한 이유로 이번 프로젝트에 적합한 모델이라고 생각되어, MobileNetV2 모델을 사용하게 되었다. 또한, MobileNetV2 모델 위에 선형 회귀와 Dropout 을 쌓아 더욱 정확도 높은 모델을 만들었다 <그림 11><그림 12>.



<그림 11>

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
mobilenetv2_1.00_224 (Model) (None, 2, 2, 1280)        2257984
_________________________________________________________________
flatten (Flatten)            (None, 5120)              0
_________________________________________________________________
dense (Dense)                (None, 1)                 5121
=================================================================
Total params: 2,263,105
Trainable params: 2,228,993
Non-trainable params: 34,112
_________________________________________________________________
```

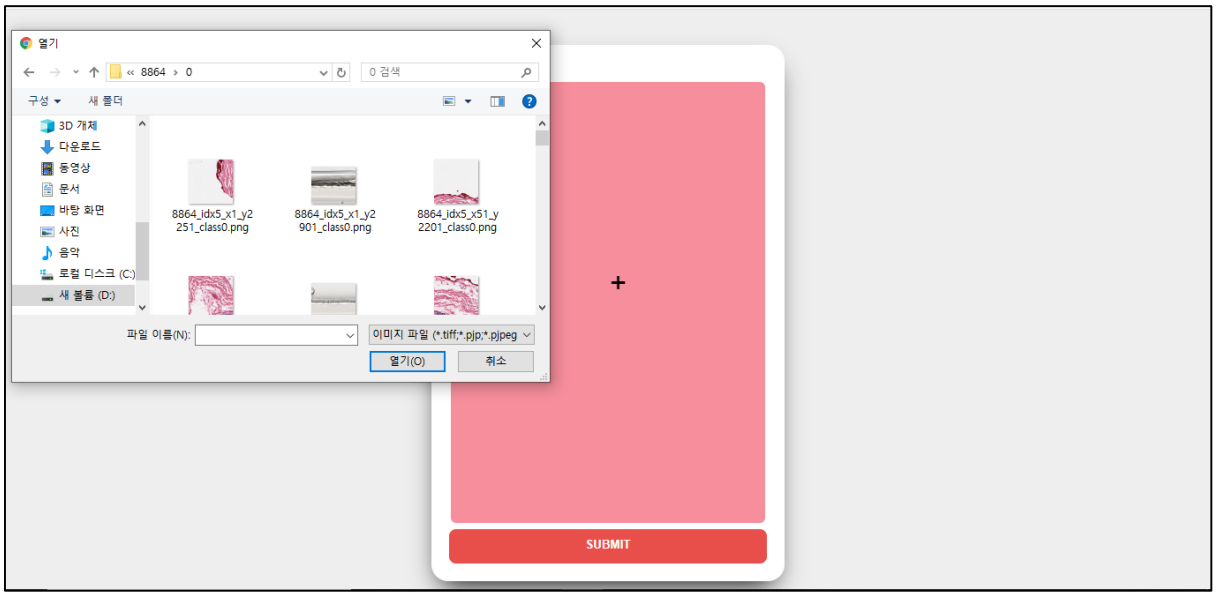
<그림 12>

주요 동작

학습시킨 예측모델을 Flask API 로 배포하여 많은 사람들이 웹을 통해서 유방암 예측이 가능하도록 하였다. 메인 화면에서 파란색 글씨를 클릭하면 이미지를 업로드 할 수 있는 페이지로 넘어가게 된다 <그림 13>.

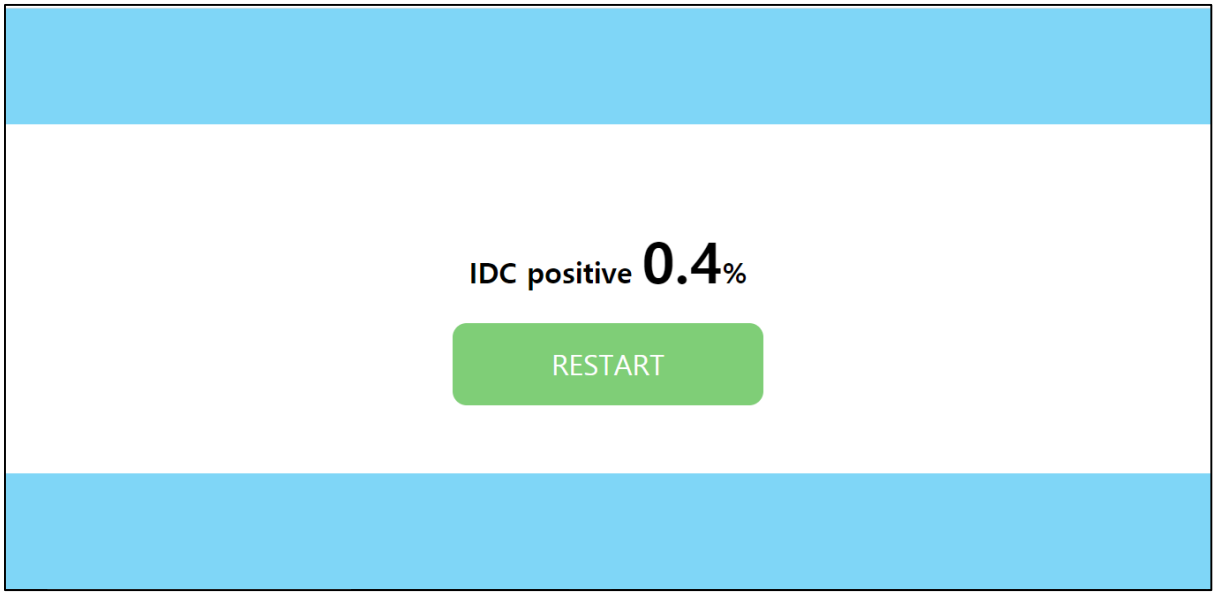
<그림 13> 메인 화면

'+'가 있는 중앙의 분홍색 사각형을 눌러주면 파일을 업로드 할 수 있는 창이 뜬다 <그림 14>.



<그림 14> 이미지 업로드 화면

업로드 한 이미지를 서버에 내장된 모델로 Predict 시켜, 양성일 확률을 출력하는 페이지이다 <그림 15>. 임의로 음성 이미지를 업로드 한 결과, 양성일 확률은 0.4%가 나온 것을 확인할 수 있다. 올바른 예측 결과가 도출되었다.



<그림 15> 결과 도출 화면

3. 기대 효과

3.1 향후 개선 사항

분석 정확도 개선

- MobileNetV2 보다 더 높은 정확도를 보장하는 다른 모델에 train data 를 학습시켜 정확도를 높일 수 있다.
- 틀린 예측을 하는 이미지를 육안으로 확인하여, 그들의 특성을 파악한다. 그 후, 그에 맞게 다양한 데이터 전처리를 시도해본다.

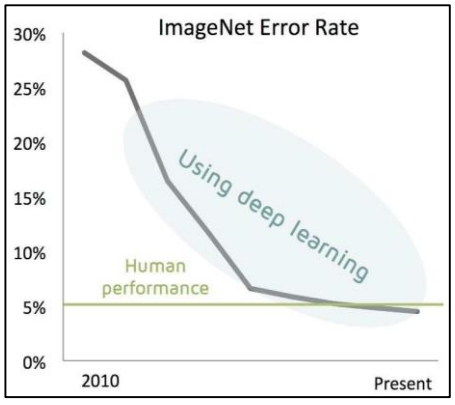
시스템 개선

- 회원가입/로그인 기능 추가해 회원제로 운영한다.
- 예측했던 모든 이미지와 예측 결과를 리스트화 하여 모든 이용자들에게 예측 사례를 공개한다.

3.2 기대 효과

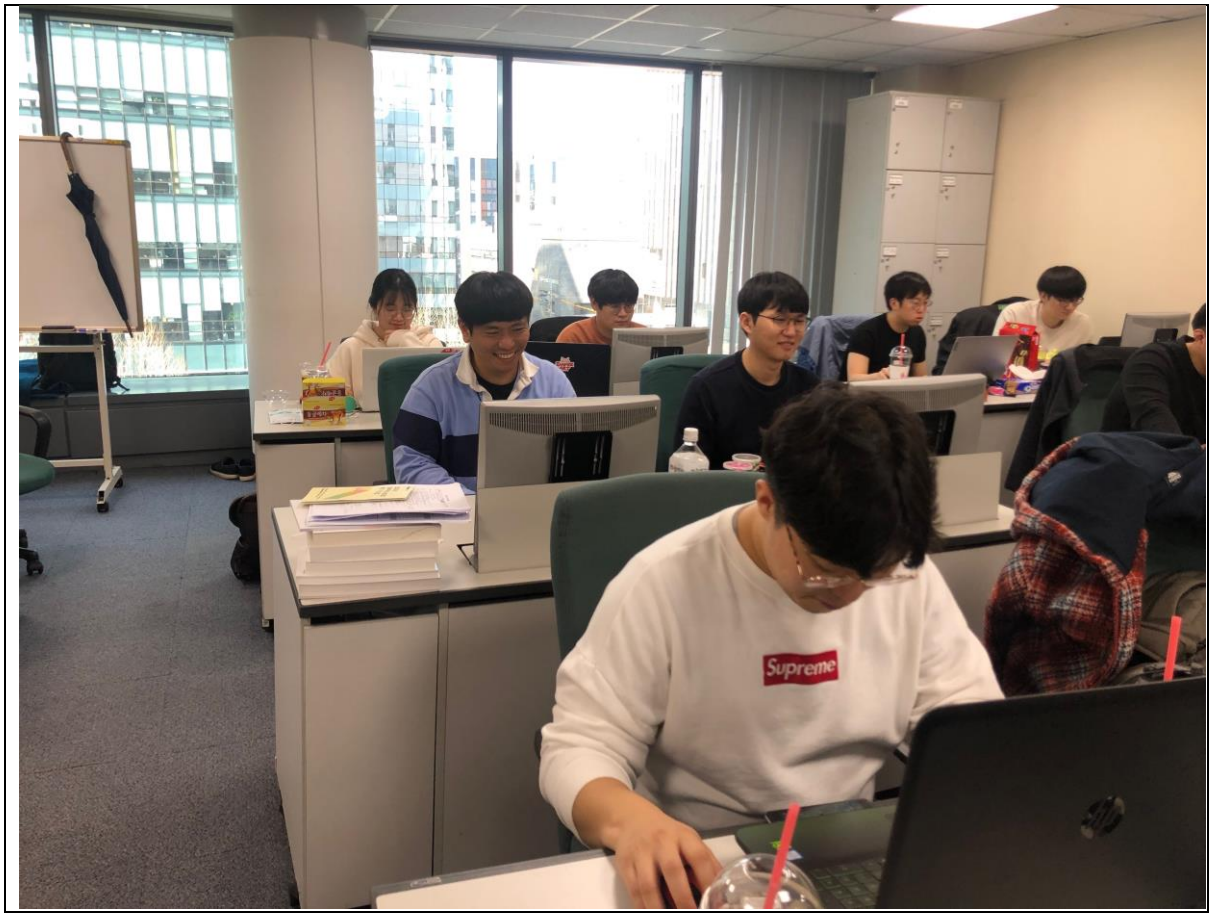
인간은 언제나 100% 정확한 판단을 할 수 없고, 어느 정도의 오차 확률이 존재한다. 그렇기 때문에 다양한 질병의 진단에 있어 의사가 잘못된 소견을 낼 가능성 또한 존재한다. 하지만 조기진단 및 조기치료가 매우 중요한 질병이 존재하고, 잘못된 진단으로 인해 조기치료 시기를 놓치면 환자의 질병은 심각한 수준의 상태가 될 수 있다.

CNN 이 점점 발전하면서 ImageNet 의 Error Rate 는 점차 줄어드는 추세이고, 현재는 인간의 눈보다 더 낮은 오차율을 보인다 <그림 16>. 즉, 의사의 판단보다 더 낮은 오차 확률로 환자의 질병 여부를 판단할 수 있다. 따라서, IDC FINDER SYSTEM 을 통해 초기에 보다 정확한 유방암 예측을 하고, 조기치료를 통한 상태 호전을 기대할 수 있다.



<그림 16>

4. 개발 후기



성명	후기
이찬호	유방암 데이터가 캐글에서 제공하는 이미지 데이터이고 프로젝트 기간이 매우 짧아 빠르지만 성능이 좋은 Keas 이미지 어플리케이션 중 MobilenetV2 를 이용해 이미지를 이진 분류로 모델을 만들게 되었다. 프로젝트를 진행하면서 세상의 모든 데이터는 캐글에서 제공하는 데이터만큼 깔끔하지 않고 가공이 되어있지 않은 데이터라는 것을 인지하게 되었으며 그에 따라 전처리 잘하는 것도 실력이라고 느끼게 되었다. 세미프로젝트 주제인 유방암에 대해 무지했지만 이번 프로젝트를 진행하면서 유방암의 종류에 대해 알게 되었으며 유방암이 여성만 걸리는 것이 아니라 남성도 걸릴 수 있다는 것에 대해 알게 되었다.
이정철	flask 를 이용한 Web Application 구성에 참여하여 웹페이지들을 서버와 연동하는 작업을 하였습니다. Back-End 개발을 처음으로 참여하였고, 개발에 처음 참여한 만큼 로그인과 CRUD 등의 다양한 기능구현들이 들어가지는 않았고, DB 연결이 이루어지지는 않았지만 웹에서 실행되는 것을 보니 뿌듯했습니다. 우리가 일반적으로 보고 있는 웹 페이지들이 보기에는 굉장히 간단하게 만들어져 있고, 왜 보기 좋게 만들지 않았을 까라는 의문이 항상 있었지만, 개발에 참여해보니 사소한 변화들임에도 Back-End 에서는 큰 변화들로 다가오는 것을 깨달았습니다.
정소현	이번 프로젝트를 통해서 교육과정 중 배웠던 딥러닝 모델인 CNN, Resnet 이외에도 이미지 분류를 수행할 수 있는 다른 많은 모델들이 존재한다는 것을 알게 되었다. 프로젝트 기간이 짧았기에 모델들 전부에 대해 완벽하게 공부를 하지 못한 것에 아쉬움이 크게 남았다. 그렇기 때문에 프로젝트가 끝나고 난 뒤에 다양한 모델들에 대해 심도 있게 공부해, 데이터와 상황에 알맞은 모델을 찾아 쓰는데 무리가 없도록 능숙해지고 싶다고 생각했다. 또한, 학부 때 정형 데이터 전처리 공부만 했기 때문에 이미지 데이터 전처리를 진행할 때 조금 어려움을 겪었다. 데이터의 종류에 따라 다양한 전처리 방법에 대해 공부하는 시간이 필요할 것 같다고 생각이 들었다.
정용주	이번 세미프로젝트에서는 프론트엔드 개발 부분을 맡아 진행하게 되었습니다. 프로젝트 주제는 유방암 디텍트 서비스로 요즘 우환 폐렴 양성 여부와 비슷한 문제라고 생각합니다. 우한 폐렴은 양성판정을 할 때 크게 2 가지 방식으로 나눠서 검사를 진행합니다. 먼저 PCR 방법을 사용하는 방법은 RNA 를 증폭하여 그 RNA 를 검출하는 방법으로 검사 시 여러 가지 에러가 발생합니다. 첫째로, 바이러스에 감염된 부위에서의 채취가 항상 선행되어야 한다는 점입니다. 만약 폐에 직접 바이러스가 침투하였다면, 폐를 점액질로 바꿔 점점 숨쉬기 힘들어지게 하고 급성 폐렴을 유발하여 이른 시일 내에 환자를 사망에 이르게 할 수 있습니다. 그러나 국가에서 시행하는 PCR 검사는 인후 부위와 코에서 바이러스를 채취합니다. 이런 경우에는 여러 번 측정해도 음성반응이 나올 수 있습니다.

	그리고 두 번째 문제로는 바이러스가 너무 미량이라 RNA 를 증폭하여도 검출이 안 되는 경우가 나올 수 있습니다. 이런 여러 가지 문제점으로 인해 PCR 검사를 하는 것보다 CT 촬영을 함으로써 검사하는 방법이 더 정확하고, 치명적인 상황의 발생을 미리 예견할 수 있는 장점이 있습니다. 그러나 의사의 판단만으로는 정확하지 않을 수 있으므로 과거의 폐렴 환자들의 양성 데이터와 음성 데이터를 학습시켜 의사의 판단을 돕고자 이번 프로젝트를 실행하여 보았습니다. 이번 프로젝트는 유방암에 대한 양성 판정이었지만, 우환 폐렴 환자의 데이터만 확보할 수 있다면, 프로젝트의 맥락은 같다고 생각합니다. 추후 데이터를 확보할 수 있게 된다면 빠른 시일 내에 우환 폐렴 양성 판정 검사 서비스를 진행할 계획입니다. 이번 프로젝트에서는 기간이 길지 않아 데이터를 가지고 딥러닝 모델을 만드는 데 주력하였으며, 프론트엔드 부문에서는 간단히 사진을 입력 받아 출력으로 진위여부만 판단하여 주면 되었기에 크게 신경 쓸 부분이 없었습니다. 파이널 프로젝트는 기간이 한 달 정도 되기 때문에 서비스할 수 있을 정도의 퀄리티를 만드는 게 목적이며, 배포하고 인터넷 커뮤니티 등에 광고하여 실제 사용해보고 사용자들의 취향에 따라 발전시켜 나가겠습니다. 이번 프로젝트에서는 기간이 길지 않아 데이터를 가지고 딥러닝 모델을 만드는데 주력하였으며, 프론트엔드 부문에서는 간단히 사진을 입력 받아 출력으로 진위여부만 판단하여 주면 되었기에 크게 신경 쓸 부분이 없었습니다. 파이널 프로젝트는 기간이 한 달 정도 되기때문에 서비스 할 수 있을 정도의 퀄리티를 만드는게 목적이며, 배포하고 인터넷 커뮤니티등에 광고하여 실제 사용해보고 유저들의 취향에 따라 발전시켜 나가겠습니다.
황지민	캐글이라는 머신러닝/딥러닝 경진대회 플랫폼을 이용해 이미지 분류를 통한 유방암 진단예측하는 세미프로젝트를 진행했습니다. 언어의 장벽을 넘어서 사람들과 의견을 공유하며 다양한 CNN 모델을 사용했습니다. 어떤 이미지에 어떤 방향으로 Convolution layer 와 activation 함수를 적용하면 좋을지 감을 익힐 수 있었습니다. 또한 다양한 시도를 통해 최근 어떻게 이미지 분류예측이 발전하고있는지 배울 수 있었고, tensorflow 와 keras 의 버전문제로 어려움이 있었지만 팀원들과 문제를 공유하면서 해결할 수 있었습니다. 또한 이러한 예측모델을 웹서버에 배포하면서 flask api 서버를 효율적으로 구축하는 방법을 터득했습니다.