“내 글씨와 닮은 글씨체는 무엇일까?”

김요은, 김성연, 김지혜

문제 정의 및 배경

- 딥러닝 프로젝트 목표: **이미지 데이터를 활용한 분류 및 예측**
  - “우리가 사용할 수 있는 이미지 데이터가 무엇일까?”
- 손글씨 데이터: 개인의 특성을 데이터화한 것 중에 체계적으로, 객관적인 수치로 만들어진 데이터
- 백인백색 다양한 글씨체 중 ‘내 글씨체와 닮은 글씨체를 추천해줄 수 있을까?’

데이터 수집

- AI_HUB “대용량 손글씨 OCR 데이터”
  - https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=605
  - 전체 연령 존재 (10대~60대)
  - 데이터 용량: 231.39 GB
  - **모델 인식률 확인용**

- 상업용 무료폰트 ‘눈누’
  - https://noonnu.cc/
  - 손글씨 12종
  - 캘리폰트 7종
  - **모델 학습용**

데이터 구성 및 전처리

데이터 분포 (대용량 손글씨 OCR 데이터)

항목	데이터량	비율(%)
10대 미만	1,275	1.2
10대	10,100	9.48
20대	30,472	28.61
30대	22,363	20.99
40대	29,863	28.04
50대 이상	12,446	11.68

- “모델에서 손글씨 인식이 잘 구현되는가?” 확인이 필요
  - “손글씨를 손글씨로 인식하는가?” (지렁이 그림으로 인식할 수도…)
  - 손글씨를 손글씨로 인식하더라도, 특정 연령의 글씨체가 제대로 인식이 되지 않아 모델 학습에 변수가 생기는 등의 문제 가능성에 대비

▶ paddle_OCR 라이브러리를 이용해 손글씨 데이터 인식 정도 확인

데이터 구성 및 전처리

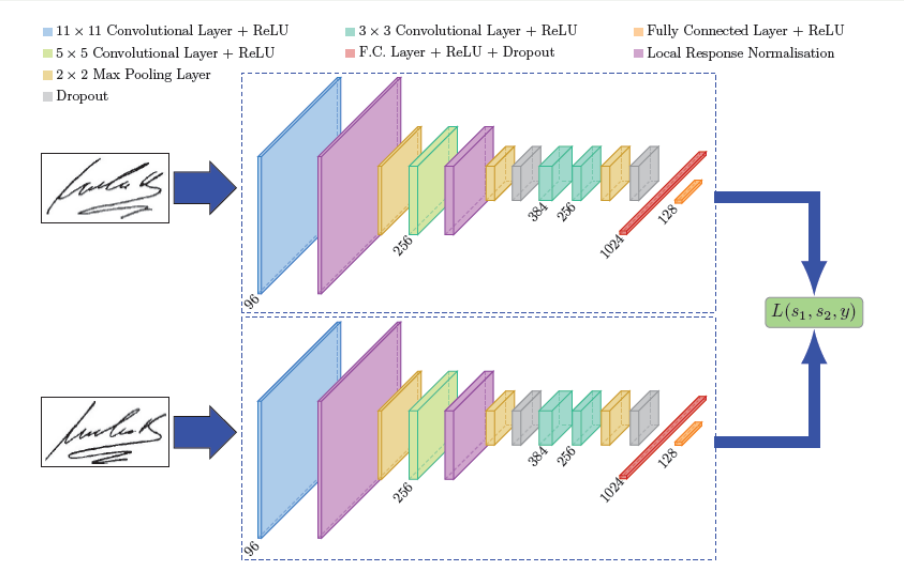
데이터 분포 (상업용 무료 폰트 데이터)

분류	폰트 이름	분류	폰트 이름
손글씨	온글잎 김콩해	캘리폰트	밀양영남루체
	잉크립퀴드체		밀양해맑은상상체
	수박양체		학교안심 가을소풍
	어비 세현체		밀양아리랑체
	인천교육자람체		태-흘림
	조군 개발새발 V5		KCC안중근체
	손편지체		
	다이어리체		
	꽃길		
	영주선비체		
	노회찬체		
	온글잎 윤우체		

- 모델을 통해 손글씨 이미지에 맞는 폰트를 찾기 위해, 입력된 손글씨의 활용 텍스트가 폰트 적용 시 출력되는 이미지를 가지고 오는 작업이 필요하였음

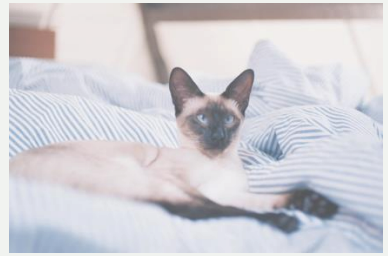
- Siamese network 모델에 폰트 이미지를 학습시키고자, 10개의 단어를 무작위로 추출하여 폰트별로 이미지를 가져와 190개의 데이터셋을 생성하였음.
  - 모델의 학습 데이터로 활용할 때 벡터화하여 사용함.

데이터 분석 및 모델링



- Siamese Network 모델 방식을 이용하여 도합 19개 폰트를 학습시킴
  - 두 개의 이미지를 동시에 입력받아 두 이미지의 특징을 추출하고, 그 유사도를 비교하는 데 매우 효과적
  - 특히, 이미지가 비슷할수록 낮은 거리(예: L2 거리)를 가지도록 학습이 가능
  - **문서 OCR**에서의 문자 비교, **서명 인증** 등에서 사용된 바 있어 폰트 이미지 비교에도 적합하여 채택

- 학습한 모델에 특정 손글씨 이미지를 주면
  해당 모델은 그 손글씨의 이미지를 추출하여 비교 후,
  해당 손글씨와 가장 비슷한 손글씨 폰트는 무엇인지
  + 그 폰트로 글씨를 쓰면 어떤지 출력함



결과

Model: "functional_1"

Layer (type)	Output Shape	Param #	Connected to
input_layer_1 (InputLayer)	(None, 224, 224, 3)	0	-
input_layer_2 (InputLayer)	(None, 224, 224, 3)	0	-
functional (Functional)	(None, 32)	30,012,384	input_layer_1[0][0], input_layer_2[0][0]
lambda (Lambda)	(None, 1)	0	functional[0][0], functional[1][0]

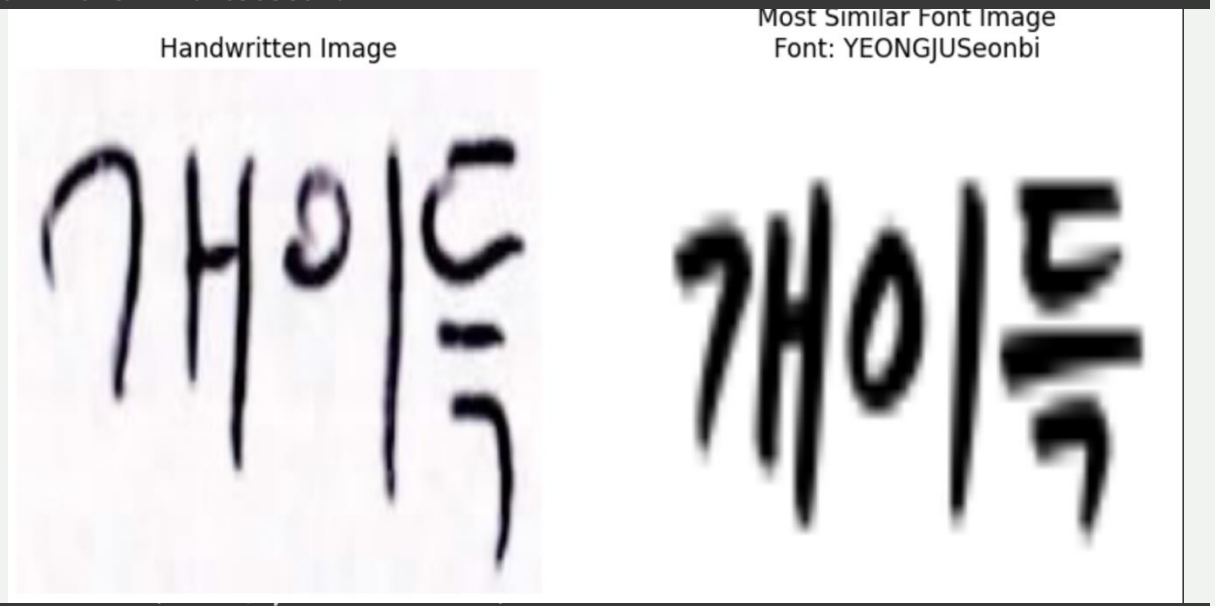
Total params: 30,012,384 (114.49 MB)
Trainable params: 29,959,264 (114.29 MB)
Non-trainable params: 53,120 (207.50 KB)

```python
# Siamese 네트워크 모델 정의
def build_siamese_model(input_shape):
    resnet = ResNet50(include_top=False, weights='imagenet', input_shape=input_shape)
    # ResNet의 출력을 플래튼(flatten)하여 Dense 레이어에 전달
    x = resnet.output
    x = Flatten()(x)
    x = Dense(64, activation='relu')(x)  # 적절한 크기의 Dense 레이어 추가
    x = Dense(32, activation='relu')(x)
```

결과

<Euclidean>

```
Most similar font image found with similarity score: 0.033076267689466476
Font name: YEONGJUSeonbi
```



<Cosine>

```
Most similar font image found with similarity score: 0.9999058842658997
Font name: 태_흘림L
```

프로젝트 의의

- “주관적인 정보의 객관적인 정보화”: 손글씨와 매우 유사한 서체를 추천
- 손글씨에서 시작하여 확장하기 좋고, 광고, 카피라이팅 등 여러 분야로 응용 및 확장이 가능할 것으로 기대

프로젝트 한계

- **주어진 상황에 대한 객관적 파악의 계기**
  - 아무리 좋은 아이디어라도 나에게 주어진 상황에 맞는 결과물을 낼 수 있던 것
  - 컴퓨팅 능력의 한계로 인해 개별 데이터셋 분석 시간이 촉박했음
- **채택된 모델 및 라이브러리의 특성에 크게 의존적**

향후 발전 방향

- 수백 가지의 손글씨 중에서 내 글씨와 비슷한 손글씨를 찾기
- “내 글씨가 인쇄체 중에서는 무엇과 가까울까?”
- 가장 최적화된 내 글씨체의 형태 찾아주기

마 무 리