

“  
어울림

모두를 위한 언어: 소통의 장벽을 넘어서.

사공명흔 엄지민 유아람 정상윤 김동휘

“

# 목차

모두를 위한 언어: 소통의 장벽을 넘어서.

---

Eowoollim

01 프로젝트 개요

04 모델

02 앱설명

05 시연

03 개발 과정

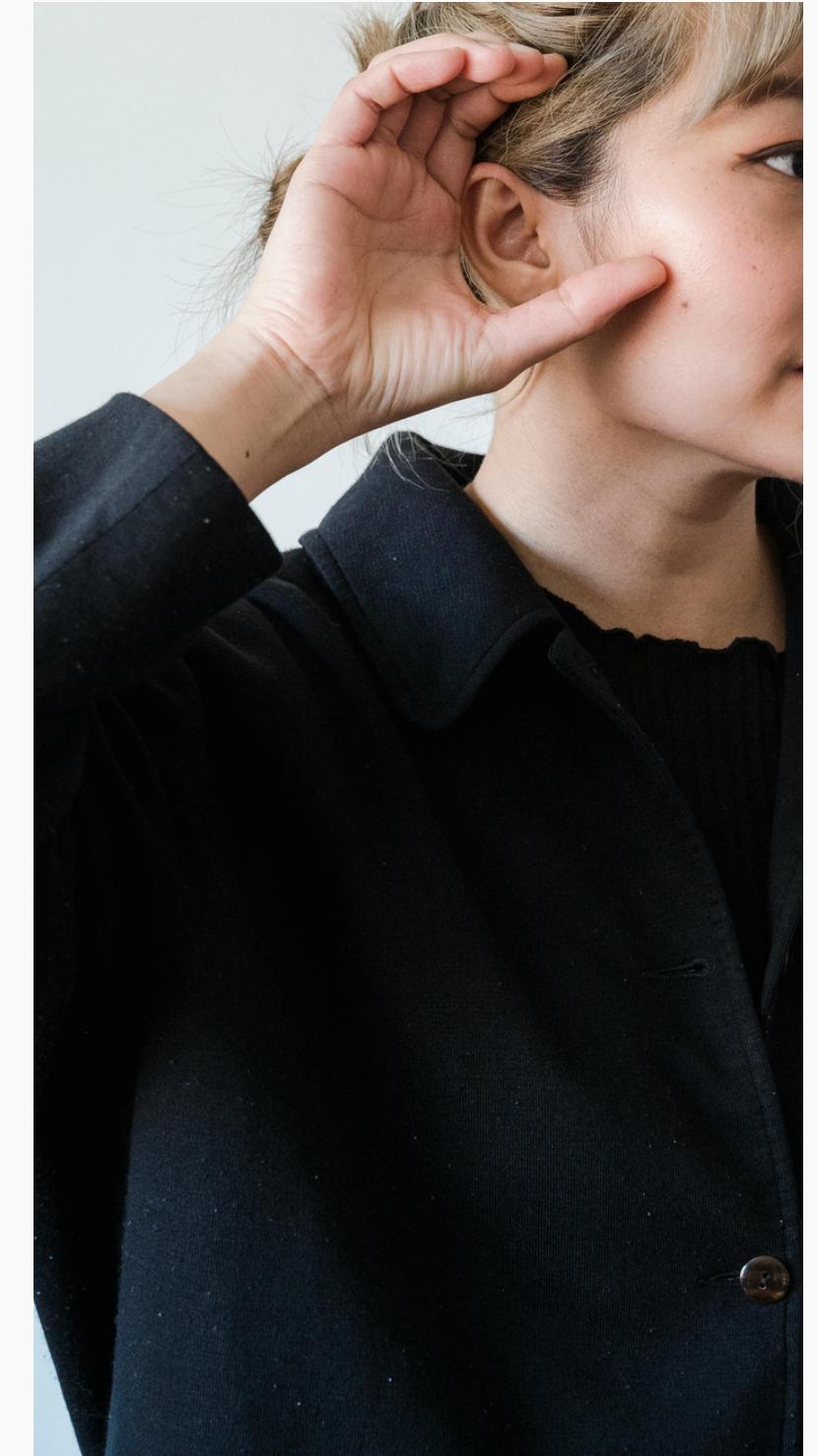
06 트러블슈팅

## 급하게 이동해야 할 때, 병원을 가야 할 때 말이 통하지 않는다면 어떨 거 같으신가요?

대부분의 사람들은 서로 소통할 때 불편함을 느낀 적이 거의 없을 거라고 생각합니다.  
하지만 서로 소통하는 것조차 힘든 사람들도 많이 있다는 것을 아시나요?

2023년 기준 전체 장애인 인구 중 청각장애인은 **16% 정도**이며 의료기술의 발달로 지체장애 등은 점점 적어지는 추세이나 그에 반해 청각장애인의 비율은 계속 증가하고 있다고 합니다.

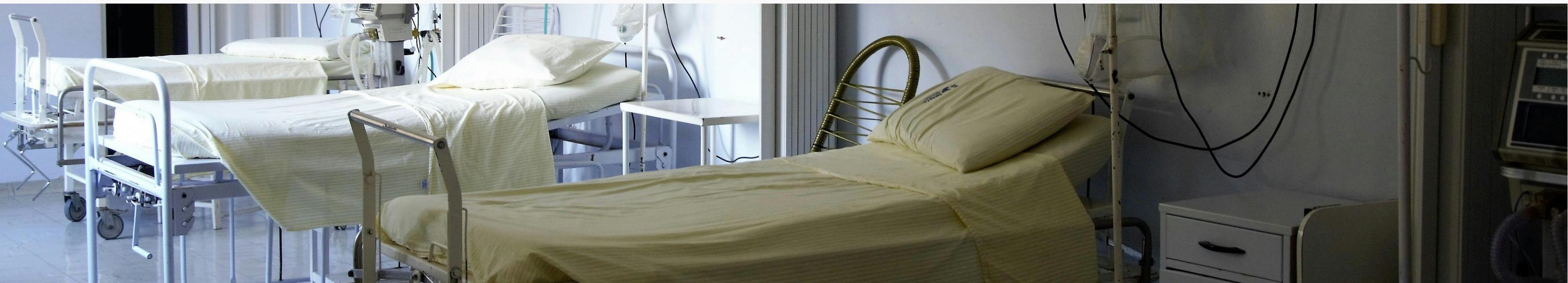
청각 장애인을 위한 환경 구성은 더 확장되어야 합니다



## 시, 청각 장애인 2명 중 1명은 의료기기 사용에 불편 느껴 (2023.12.01)

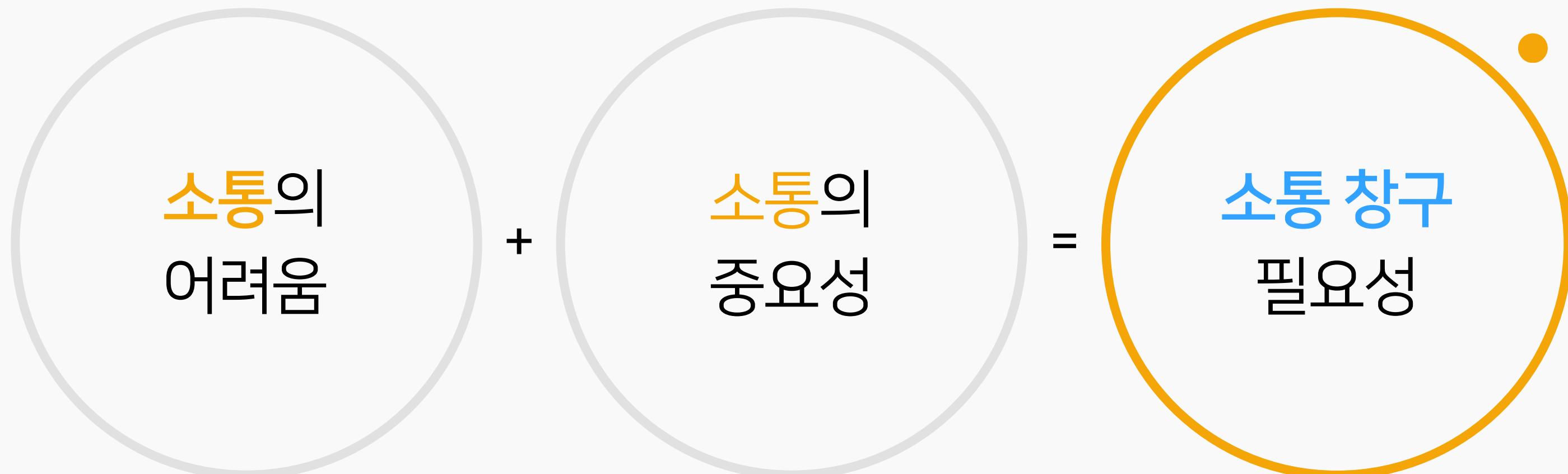
설문은 장애인 단체와 협력 아래, 시각장애인 44명·청각장애인 69명 총 113명을 대상으로 지난 9월 1일부터 10월 31일까지 진행됐다.

설문 결과, 의료기기 사용 시 불편함을 느낀다고 답한 사람은 전체의 46.9%(53명)로 나타났다. '보통이다'라고 답변한 비율은 19.5%(22명), '그렇지 않다'고 답변한 비율은 28.3%(32명)였다. 이들은 주로 전원 버튼 위치나 버튼별 기능 구분 등 의료기기 사용 정보를 확인하는 일에 어려움을 호소했으며, 의료기기 사용 시 주의사항·유효기간 등 정보 또한 확인이 어렵다고 답했다.



청인에게는 쉬운 소통  
하지만 농인에게는 쉽지 않습니다.

소통은 사람 간의 기본 활동이자 가장 중요한 활동입니다.  
그러나 농인에게 **소통은 너무나도 큰 벽입니다.**



## 어울림이란?

청인과 농인의 편리한 소통을 위한 프로그램으로 농인의 수화는 청인에게 음성으로 출력됩니다.



AI허브에 있는 수어 데이터에는  
두가지의 학습 데이터가 존재합니다.

01

영상

mp4 확장자의 다양한 영상 데이터

02

키포인트

json 확장자의 다양한 데이터

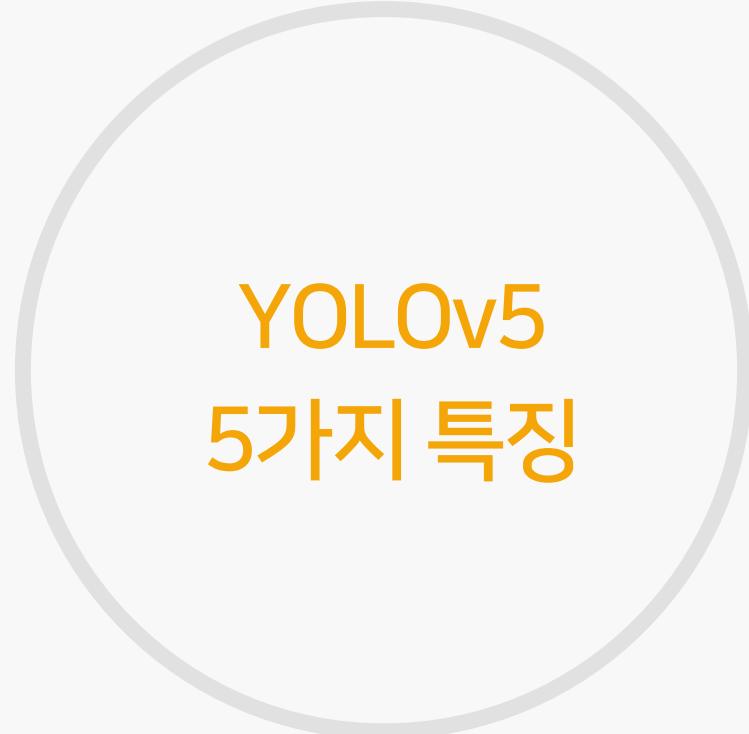
왜 두가지 다 학습을 진행했을까요?

영상 데이터, 키포인트 데이터 둘 중 어떤 데이터로 학습을 시켰을 때  
더 좋은 성능을 낼 수 있는지를 확실하게 알지 못해 좋은 성능을 내고자

3팀 내에서 영상팀, 키포인트 팀을 나눠  
두 가지의 데이터로 다양한 방법을 이용하여 학습을 진행하였습니다.

## YOLO 모델을 이용한 동영상 학습방법

YOLO 모델이 학습한 이미지 속 각각의 아키텍처들을 구분하는 성능이 뛰어나다는 점을 이용하여, 각 수어를 나타내는 이미지들의 손 모양이 다르므로 특정 손 모양이 나왔을 때 해당하는 뜻을 출력할 수 있지 않을까 하는 생각이 들어 최종적으로 선택하였습니다.



YOLOv5  
5가지 특징

01. 높은 정확도

02. 빠른 속도

03. 간편한 사용

04. 다양한 애플리케이션에 적용

05. 가벼운 모델

## 원활한 학습을 위한 데이터 전처리 작업 진행

images

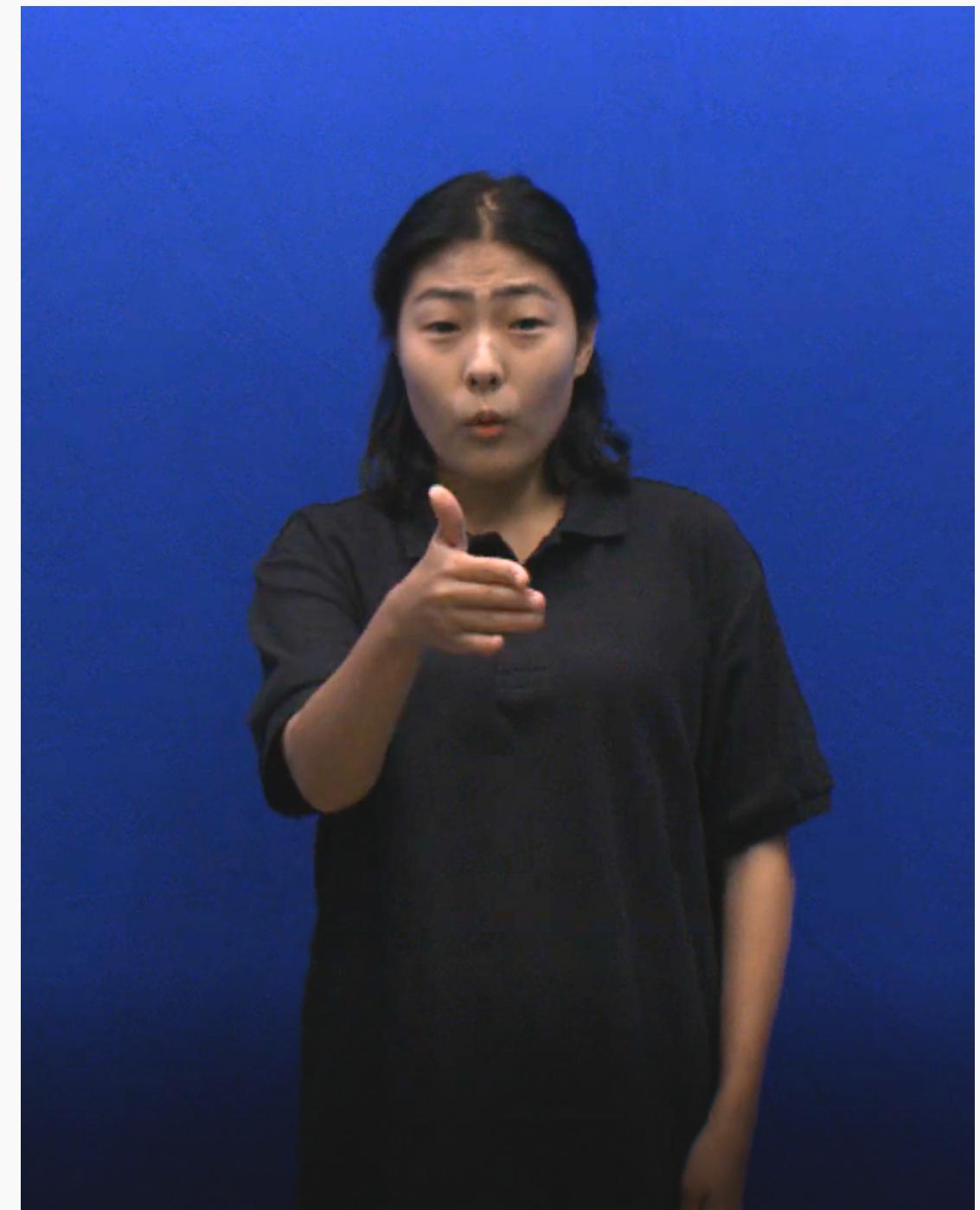
aihub에서 받은 수어 영상들 중 차렷 자세를 제외한 수어 부분만 최대한 추려내어  
프레임 단위로 쪼갠 이미지로 변환하여 이미지 데이터 준비.

labels

morepheme(json)파일 속 단어들의 뜻을 딕셔너리 형태로 바꾼 후 단어의 번호에 맞는  
클래스 번호를 할당. (예: WORD0001 : 클래스번호:0, 뜻:고민)

위 과정으로 만들어진 데이터들을 활용하여 **yaml 파일 생성**

<영상 캡쳐본>

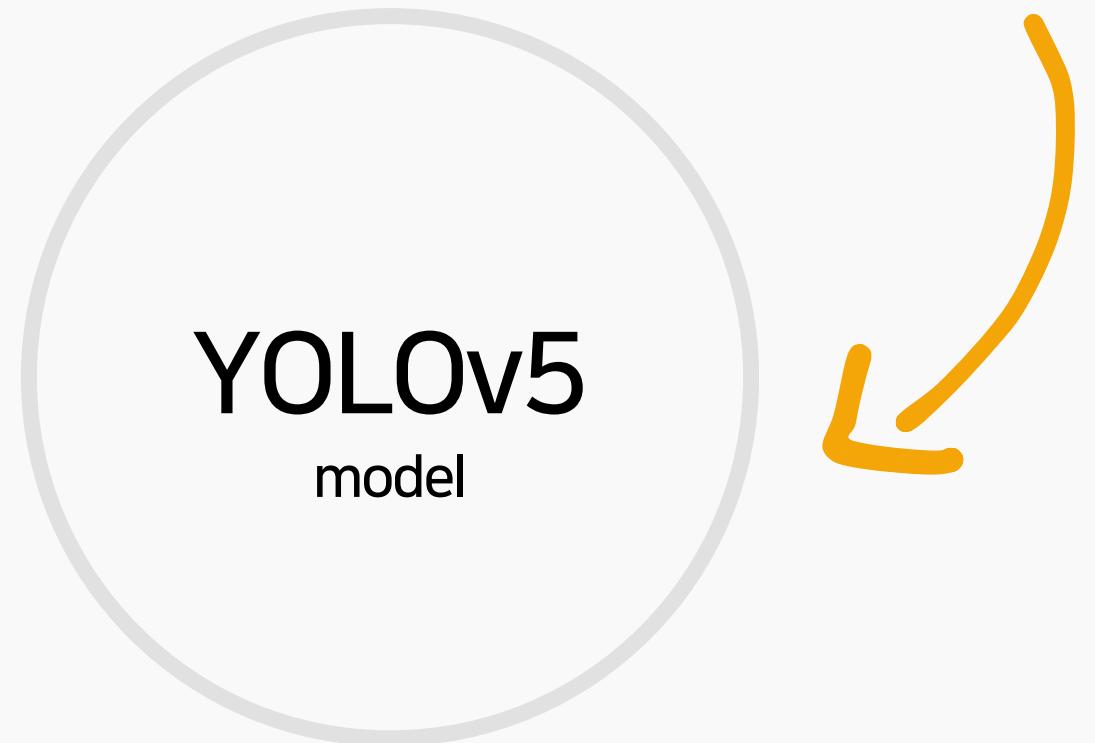


## 전처리한 데이터를 YOLO 모델을 이용하여 학습

입력

입력 이미지의 크기 : 640X640으로만 진행

가중치: [yolov5s.pt](#), [yolov5s.pt](#) [yolov5l.pt](#)를 사용



결과

### yolov5s.pt

테스트 과정에서 사용. 단어 3개, 사람 2명 분의 데이터를 학습시켰을 때 confidence threshold가 0.2 일 때도 각 단어를 잘 인식하는 모습을 보였으나, 이후 단어 10개, 사람 3명 분의 데이터를 학습시켰을 때 confidence threshold가 0.01 정도 됐을 때 **4개의 단어를 인식하는 수준으로 정확도가 현저히 떨어지게 됨.**

### yolov5l.pt

이후 단어 10개, 사람 3명 분의 데이터를 학습시킨 후 confidence threshold가 0.05 정도 됐을 때 7개의 단어를 인식하는 수준으로 yolov5s.pt를 가중치로 학습시켰을 때보다 **확실히 정확도가 올라간 모습을 보였습니다.** 하지만 모델의 용량이 yolov5s에 비해 현저히 커지고 무거워지는 부분을 확인

## 영상학습 최종모델

yolov5s보다 yolov5l이 더 나은 성능을 보인 부분과 yolov5l을 사용하여 만든 모델이 상당히 무겁다는 단점을 생각하여 최종 학습 시엔 yolov5m.pt를 사용하여 단어(클래스) 100개, 사람 13명 분의 데이터를 학습시켰습니다.

클래스가 늘어날수록 정확도가 현저히 떨어지는 모습을 보여주었습니다.



클래스 개수를 100개로 진행한 최종 학습 후의 결과 yolov5s.pt보다 정확도가 높은 yolov5m.pt를 사용했음에도 confidence threshold가 0.001 일 때 마저도 뜻을 제대로 번역해내지 못하는 모습을 발견하였습니다.

차렷 자세에서 엉뚱한 뜻이 출력되는 경우가 많았던 점을 보아 이미지 데이터를 최대한 정제했음에도 불구하고 차렷자세의 이미지 데이터가 남아 있었던 것으로 보이며, 데이터의 양과 질 또한 중요해 보인다는 결론을 도출해냈습니다.

## 시공간지도사상기법을 이용한 키포인트 학습

위 논문을 참고하여 수어 영상을 시공간지도로 만든 후  
이미지를 CNN 전이모델에 학습을 진행하는 방법으로 학습

참고논문

## 신체 키포인트의 시공간 지도 사상 기법을 사용한 수어 번역 모델

[https://kw.dcollection.net/public\\_resource/pdf/200000651674\\_20240306151530.pdf](https://kw.dcollection.net/public_resource/pdf/200000651674_20240306151530.pdf)



각 프레임 별 신체 키 포인트의 x축, y축, 정확도를 정규화 하여  
각각을 색상 좌표로 변환하여 픽셀에 매핑

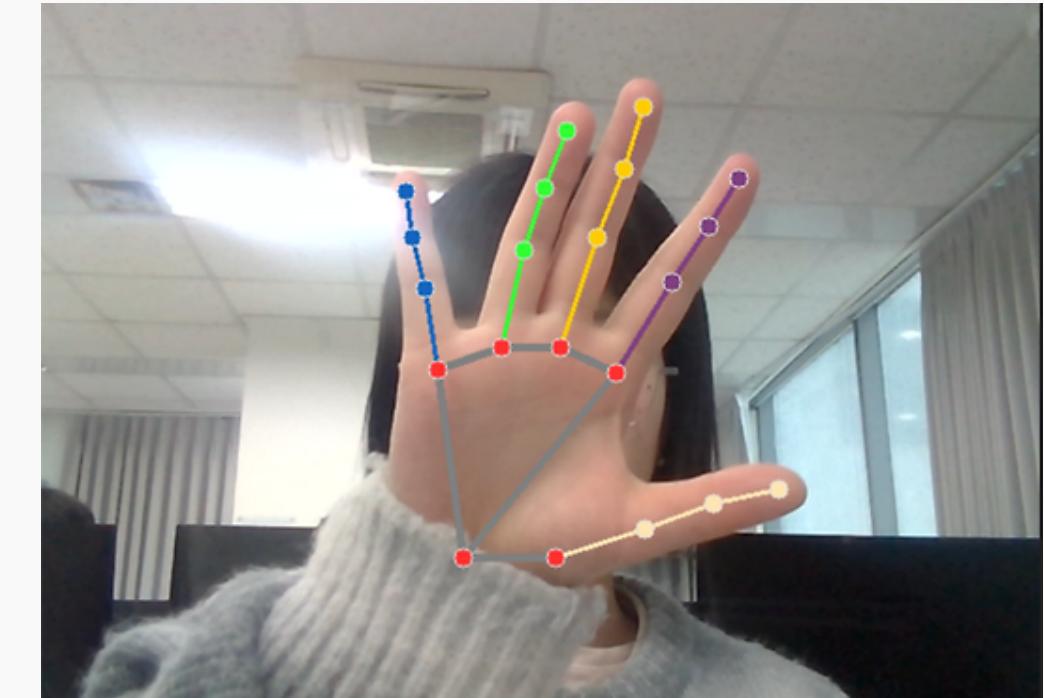
## mediapipe를 이용한 keypoint 추출

수어라는 주제 특성 상 정지된 동작으로 단어를 추측하는게 아닌 수어를 하는 과정 또한 중요하기 때문에 키포인트팀 모델은 정확한 keypoint를 뽑아 앞에서 말한 시공간지도사상기법을 이용하여 시공간지도를 저장하는것이 중요한 포인트입니다.

그래서 키포인트팀은 mediapipe란 구글에서 만든 모션감지 모델로 손의 키포인트를 출력하였습니다.



### mediapipe 웹캠 적용 예시



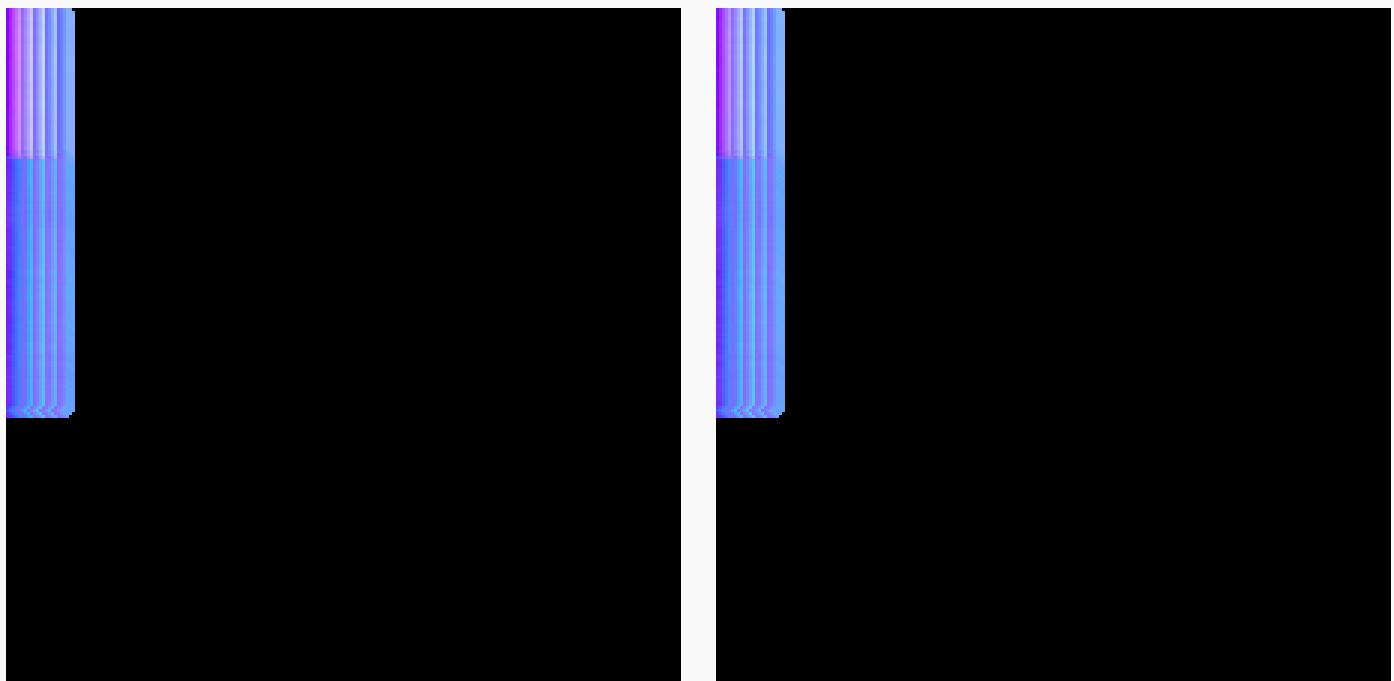
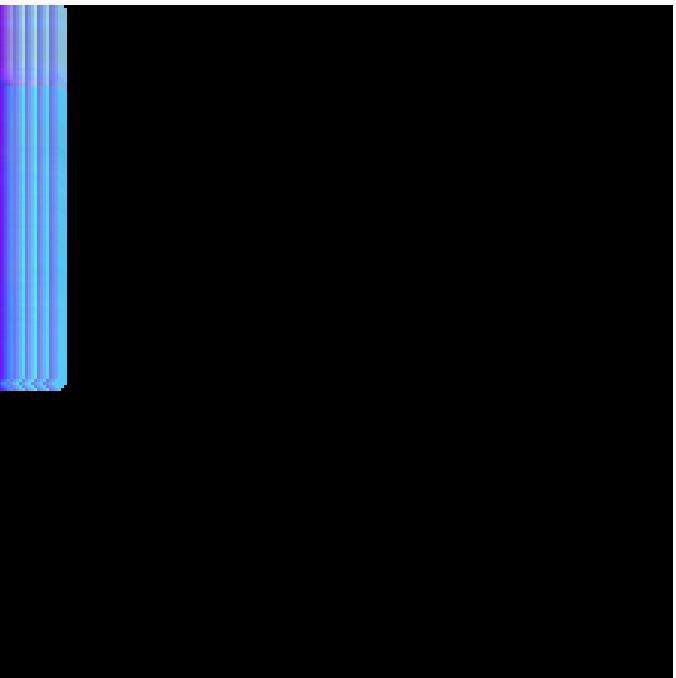
좌표: 0.4500951/009077754, Y 좌표: 0.9281275808415833, Z 좌표:  
좌표: 0.41272565722465515, Y 좌표: 0.8119428157806396, Z 좌표:  
좌표: 0.3857714831829071, Y 좌표: 0.7359880805015564, Z 좌표:  
좌표: 0.35577258467674255, Y 좌표: 0.6802708506584167, Z 좌표:  
좌표: 0.4779389202594757, Y 좌표: 0.9230219125747681, Z 좌표:  
좌표: 0.4352216422557831, Y 좌표: 0.7971619963645935, Z 좌표:  
좌표: 0.4033065438270569, Y 좌표: 0.7153002619743347, Z 좌표:  
좌표: 0.3728211522102356, Y 좌표: 0.6462274789810181, Z 좌표:  
좌표: 0.517368733882904, Y 좌표: 0.9238245487213135, Z 좌표: -  
좌표: 0.4773833155632019, Y 좌표: 0.8036816120147705, Z 좌표:  
좌표: 0.43631863594055176, Y 좌표: 0.7294983267784119, Z 좌표:  
좌표: 0.3965473771095276, Y 좌표: 0.6693732738494873, Z 좌표:  
좌표: 0.5683152675628662, Y 좌표: 0.92885822057724, Z 좌표: -0  
좌표: 0.5663760304450989, Y 좌표: 0.8207968473434448, Z 좌표:  
좌표: 0.5531829595565796, Y 좌표: 0.7506706714630127, Z 좌표:  
좌표: 0.5341747999191284, Y 좌표: 0.6926537752151489, Z 좌표:

## keypoint 정규화 후 컬러 이미지로 변환

보통 AI허브에서 다운받은 자료들은 깔끔하게 정제되어있는 데이터들이 많습니다.  
하지만 실제 사용할때는 학습할때 넣은 데이터들과는 다릅니다.  
사람이 왼쪽에 있을수도 있고 오른쪽에도 있을수가 있습니다.

이렇게 되면 좌표값이 전부 달라져 제대로된 결과가 나오지않아  
**정규화를 해주어 모든 키포인트를 0~1 사이의 값으로 만들어주면 모든 영상에서**  
비슷한 값을 뽑아낼 수 있습니다.

정규화 진행 후 시공간지도 변환



## CNN 모델을 커스텀하여 전이학습 진행

이미지를 224 X 224 사이즈로 통일화시켜 모델에 입력

ResNetCustom Class 객체를 만들어 학습을 진행하였고 데이터의 전처리의 시간이 길지만 모델의 학습 시간은 비디오를 직접 학습시키는 것보다 현저히 짧아 전처리와 모델의 학습을 병렬적으로 진행하여 점진적으로 모델의 성능을 향상시킬 수 있다는 장점이 있습니다.

또한 기존 모델은 output이 1000개 이하인 경우에 사용되는 모델이기 때문에 3000개의 단어에서 많이 쓰이는 단어를 정제한 후 982개의 단어를 학습시켰습니다.

```
class ResNetCustom(nn.Module):
    def __init__(self, num_classes=1000):
        super(ResNetCustom, self).__init__()
        self.resnet = models.resnet50(pretrained=True)
        self.resnet.fc = nn.Linear(2048, num_classes)

    def forward(self, x):
        x = self.resnet.conv1(x)
        x = self.resnet.bn1(x)
        x = self.resnet.relu(x)
        x = self.resnet.maxpool(x)

        x = self.resnet.layer1(x)
        x = self.resnet.layer2(x)
        x = self.resnet.layer3(x)
        x = self.resnet.layer4(x)

        x = self.resnet.avgpool(x)
        x = torch.flatten(x, 1)
        x = self.resnet.fc(x)

    return x

# 모델 생성
model = ResNetCustom().to(device)

# 모델 출력
print(model)
```

## 어울림의 최종 모델은?

keypoint

동영상을 학습시킨 YOLO 모델과 키포인트를 학습시킨 ResNet 모델을  
비교해본 결과 키포인트를 학습시킨 모델의 정확도가 더 높아 키포인트팀의 모델로 최종 결정하였습니다.

## 모델로 뽑은 단어를 자연어 처리 모델에서 처리

수어와 한국어는 문법이 많이 다릅니다.

예를들어 수어로 “나폴레옹 나라를 정복하다 정복하다 정복하다”라는 문장이 있다고 가정했을 때 이를 한국어로 번역하면 “**나폴레옹은 여러 나라를 정복했다**”라는 뜻이 됩니다.

그래서 어울림팀은 ResNet모델로 뽑은 단어들을 GPT API에 넣어 자연어 처리를 진행 하여 자연스러운 문장으로 변환해줍니다.



모두를 위한 언어: 어울림  
앱 시연이 있겠습니다

## 영상팀 학습 YOLO 모델 트러블 슈팅

### 문제

초기에 단어 3개만 학습시켰을 때는 학습 후 테스트 결과 학습한 수어를 잘 구별해 냈으나, 학습되는 단어(클래스)가 많아질수록 학습 결과가 정확도(P, mAP50, mAP50-95 등으로 표현된 수치)가 0.7 이상으로 어느정도 준수한 수준으로 나옴에도 불구하고 테스트 결과는 좋지 못함.

### 원인

YOLOv5와 같은 객체 감지 모델의 특성상 클래스가 많아질 수록 각 클래스 간의 미묘한 차이를 구별해내는 것에 어려움을 느낌.

### 해결

클래스의 수를 3000(기존목표) -> 100 -> 10 순으로 줄여가며 단어를 정확히 찾을 수 있는 범위를 파악하면서 학습

## 키포인트 학습 트러블 슈팅

### 문제

기존의 계획은 AI Hub에서 제공하는 keypoint를 사용해 모델을 학습시키고 결과값을 예측하려고 했으나, 동일한 영상에 대해 mediapipe로 뽑은 keypoint와 Hub에서 제공하는 keypoint가 다르고, 프레임별로 keypoint가 존재하기 때문에 데이터의 양이 방대하여 학습에 많은 시간이 소요

### 해결

keypoint의 형식을 맞추기 위해 주어진 keypoint를 사용하지 않고 직접 원천 데이터에서 keypoint를 추출하고, 프레임 별로 주어지는 point 값을 압축하기 위해 각 keypoint의 x, y, score 값을 픽셀에 맵핑하여 하나의 이미지에 전체 프레임에 대한 정보를 담아 이를 학습시킴

## 영상 전처리 트러블 슈팅

### 문제

본래 계획은 사용자의 별다른 노력 없이 실시간으로 수어 영상을 채집하여 이를 문장으로 번역하려고 하였으나, 수어와 자연어의 형태와 기능이 달라 자연어 처리를 해야하는 과정에서 어느 부분이 문장의 끝인지와, 단어를 나타내는 수어의 끝이 필요하게 됨

### 해결

버튼을 사용하여 사용자가 직접 문장의 시작과 끝을 정하도록 하고 이를 문장 단위로 하여 서버로 송신한 뒤 단어를 뽑아 자연어 처리를 진행 또한, 단어와 단어 사이는 사용자의 손을 프레임 밖으로 뺏으로써 mediapipe가 손을 인식하지 못할 때를 하나의 단어로 인지

## 리액트 네이티브 Expo 트러블 슈팅

문제

리액트 EXPO를 사용해 동영상을 촬영하고 이를 파이썬 서버로 전송하는 과정에서 파일의 경로는 제대로 가져오지만 이를 서버로 보내지는 못하는 오류가 발생

원인

리액트에서 Flask로 파일을 전송하려면 FormData 객체에 보낼 데이터를 저장하여 보내야 했음

해결

FormData객체를 생성하고 content-type을 multipart/formdata로 설정, 객체에 파일 이름과 형식, 파일의 내용을 담아 서버로 전송