Sign-to-Speech Model for Sign Language Understanding: A Case Study of Nigerian Sign Language

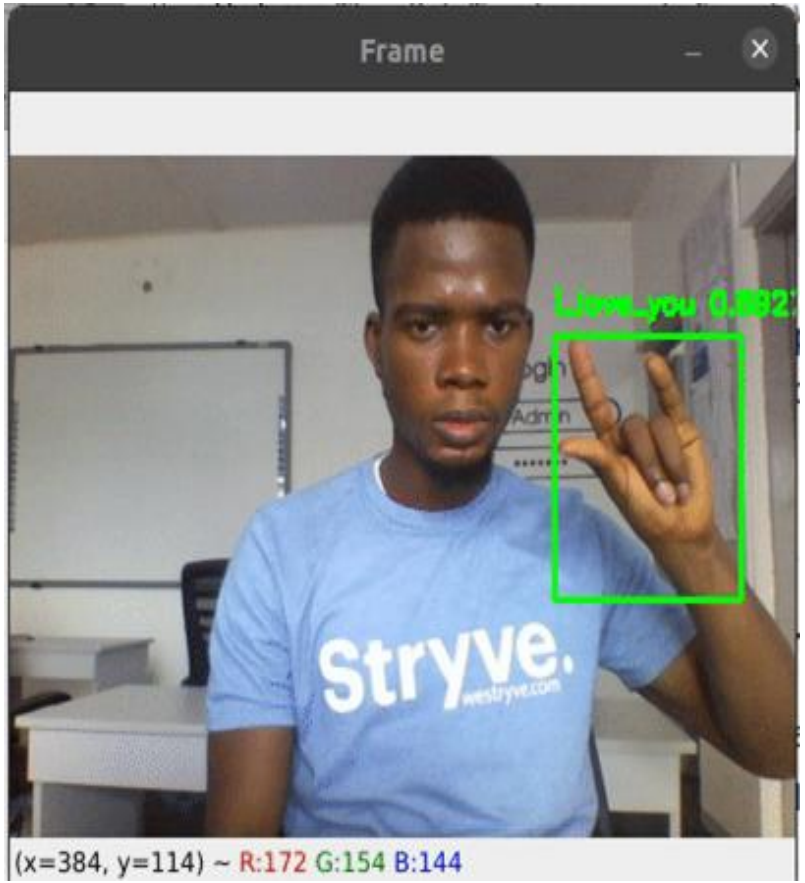
Steven Kolawole, Opeyemi Osakuade, Nayan Saxena, Babatunde K. Olorisade

Presenting at *NCS AI Summit '22*

March 24, 2022

Started as a fun project...

An end-to-end working prototype that detects sign language meanings in images/videos and generate equivalent, realistic voice of words communicated by the sign language, in real-time.



Collaborators/Mentors

Researcher, Data Science Nigeria

Senior Lecturer, CardiffMet

Maths & Stats student, UofT

ML Collective

Paper got accepted into a NeurIPS Workshop (ML4D)!

Sign-to-Speech Model for Sign Language Understanding: A Case Study of Nigerian Sign Language

Steven Kolawole
Federal University of Agriculture, Abeokuta
ML Collective
kolawolesteven99@gmail.com

Opeyemi Osakuade
Data Science Nigeria
osakuade@datasciencenigeria.ai

Nayan Saxena
University of Toronto
ML Collective
nayan.saxena@mail.utoronto.ca

Babatunde Kazeem Olorisade
Cardiff Metropolitan University
kolorisade@cardiffmet.ac.uk

Abstract

Through this paper we seek to reduce the communication barrier between the hearing-impaired community and the larger society who are usually not familiar with sign language in the sub-Saharan region of Africa with the largest occurrences of hearing disability cases, while using Nigeria as a case study. The dataset is a pioneer dataset for the Nigerian Sign Language and was created in collaboration with relevant stakeholders. We pre-processed the data in readiness for two different object detection models and a classification model and employed diverse evaluation

Talk Overview

1. Background

1. Dataset

1. Experiments, Evaluation and Deployment

1. Demo

Background

Motivation

To reduce the communication barrier between the hearing impaired community and the larger society with a focus on sub-Saharan Africa, which is the region with most cases of hearing disabilities by developing a lightweight sign-to-speech solution that works in real-time.

Motivation

Lack of solutions for this problem in sub-Saharan Africa is mostly due to two factors:

i. The sign language data in the region is low resourced

ii. There are increasing complexities and advanced tools required to deploy these solutions in real-life environments.

Motivation

Lack of solutions for this problem in sub-Saharan Africa is mostly due to two factors:

i. The sign language data in the region is low resourced

Motivation

Lack of solutions for this problem in sub-Saharan Africa is mostly due to two factors:

i. The sign language data in the region is low resourced

ii. There are increasing complexities and advanced tools required to deploy these solutions in real-life environments.

Approach

Approach

- Create a novel dataset for a sub-Saharan country sign language

  (using Nigerian Sign Language as a case study)

Approach

- Create a novel dataset for a sub-Saharan country sign language

  (using Nigerian Sign Language as a case study)

- Sign-to-text conversion experiments with several models

Approach

- Create a novel dataset for a sub-Saharan country sign language

  (using Nigerian Sign Language as a case study)

- Sign-to-text conversion experiments with several models
- Deploy the best-performing model for real-time usage while converting sign to speech

Dataset

Dataset Creation

Over 8000 images created by;

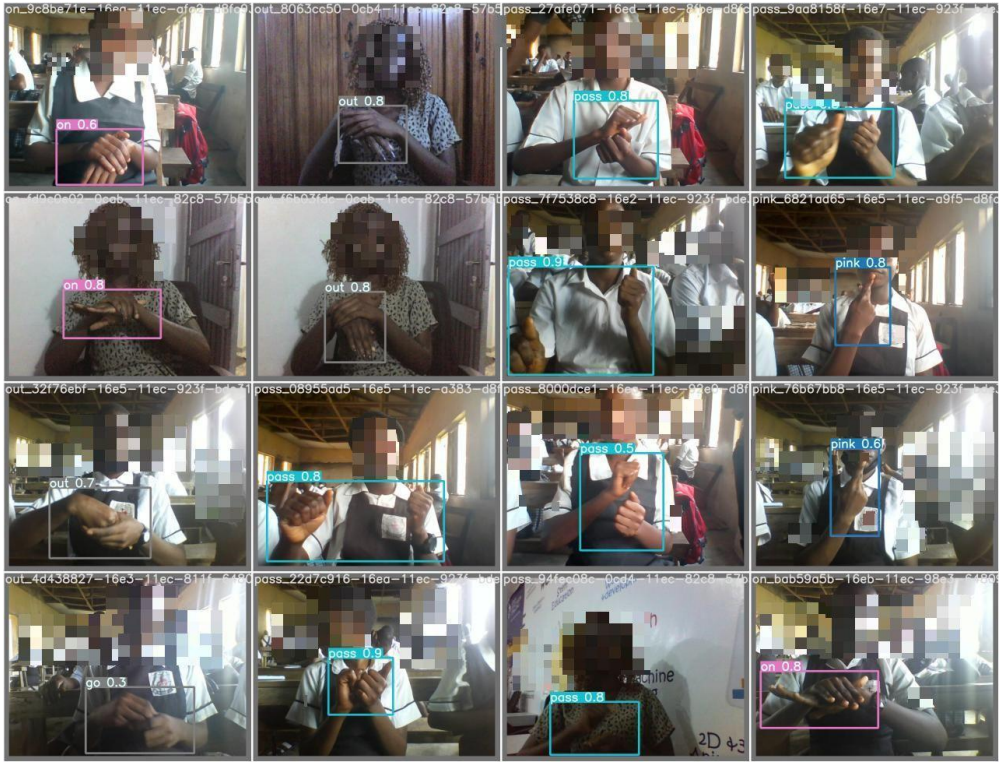
- A TV sign language broadcaster
- 20 teachers and students from 2 special education schools in Nigeria

ω

θ

Dataset Creation

A widely-dispersed dataset of 20+ individuals captured in diverse backgrounds and lighting conditions.



Dataset Preprocessing & Annotation

- Data cleaning reduced dataset to 5000 images;
- Using LabelImg, images were annotated for Object Detection in both TXT and XML formats.

ω

θ

Experiments, Evaluation and Deployment

Models

- YOLOv5 model
- Single-Shot Detector using Resnet50 FPN
- Classification model using MobileNetv2

ω

θ

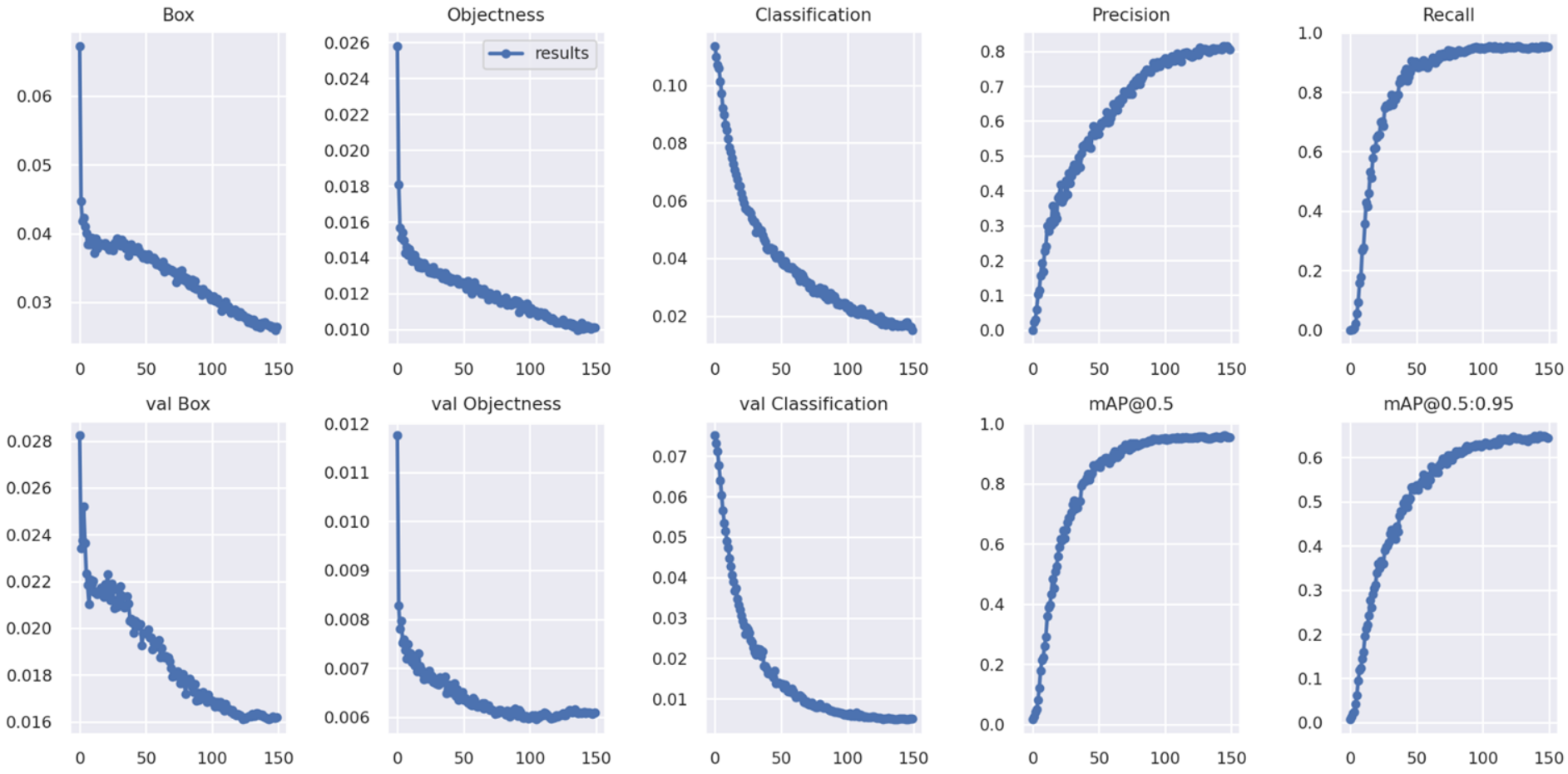
YOLO model

- TXT-format annotations.
- HSV manipulations, Scaling, and LR-flipping applied.
- Training over 150 epochs in batch size of 16.

ω

θ

YOLO model's Performance



SSD model

- TFOD API to access and finetune SSD ResNet50 V1 FPN model,
- ResNet50 was pre-trained on the COCO 2017 dataset.
- Images and Annotations converted to TFRecord format.
- Only H-flipping and cropping applied and training across 40,000 train steps.

ω

θ

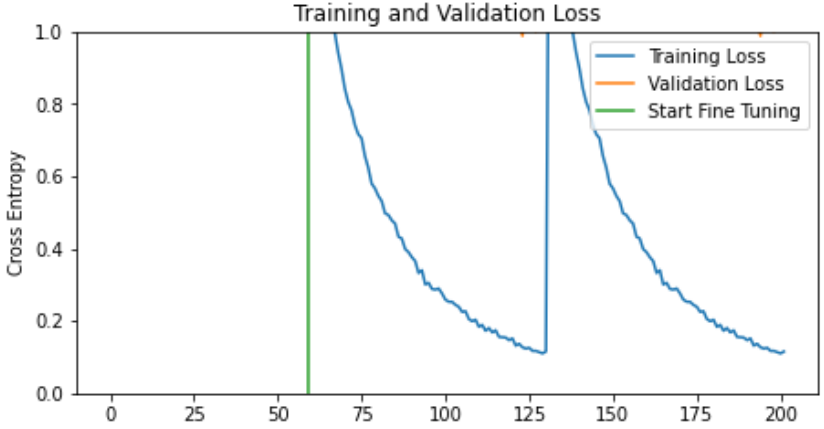
Classification Model

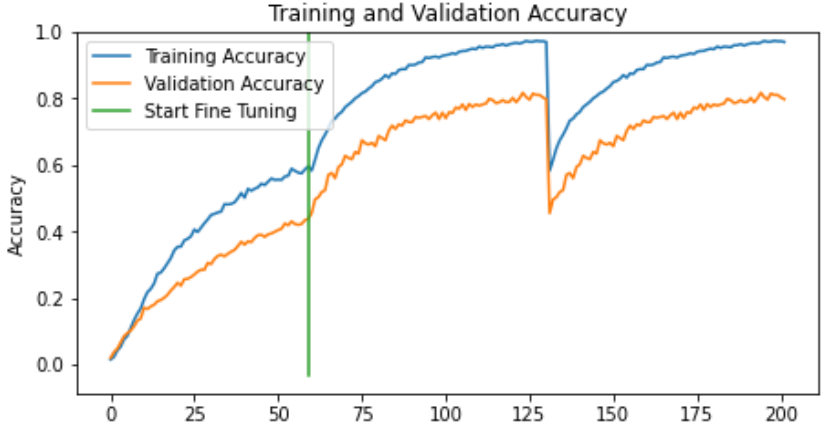
- `MobileNetv2 as our Classifier`

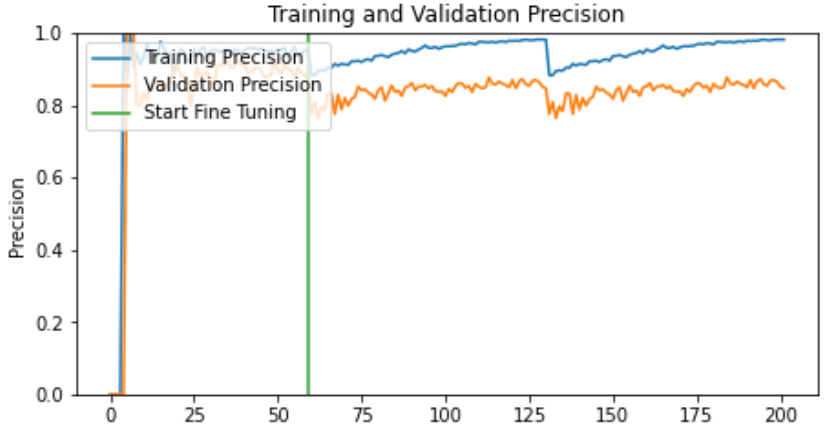
- 60 epochs of training using feature extraction only
- 140 epochs of training after fine-tuning the model

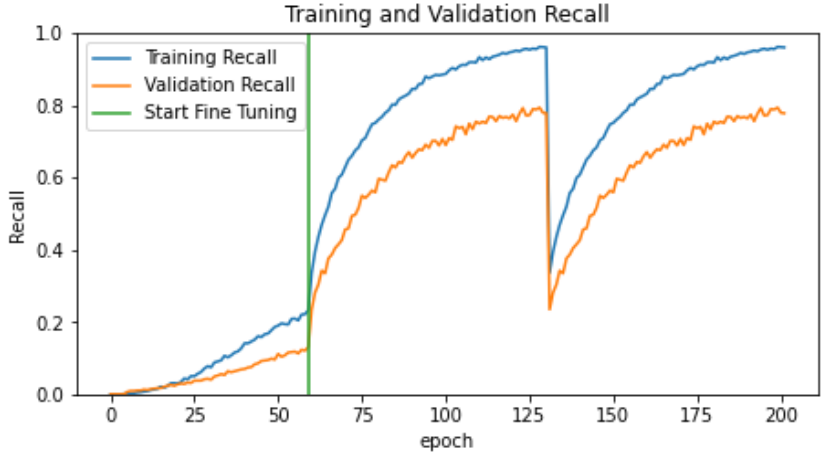
θ ω

Classifier's Performance: Feature Extraction vs Fine-tuning









Evaluation

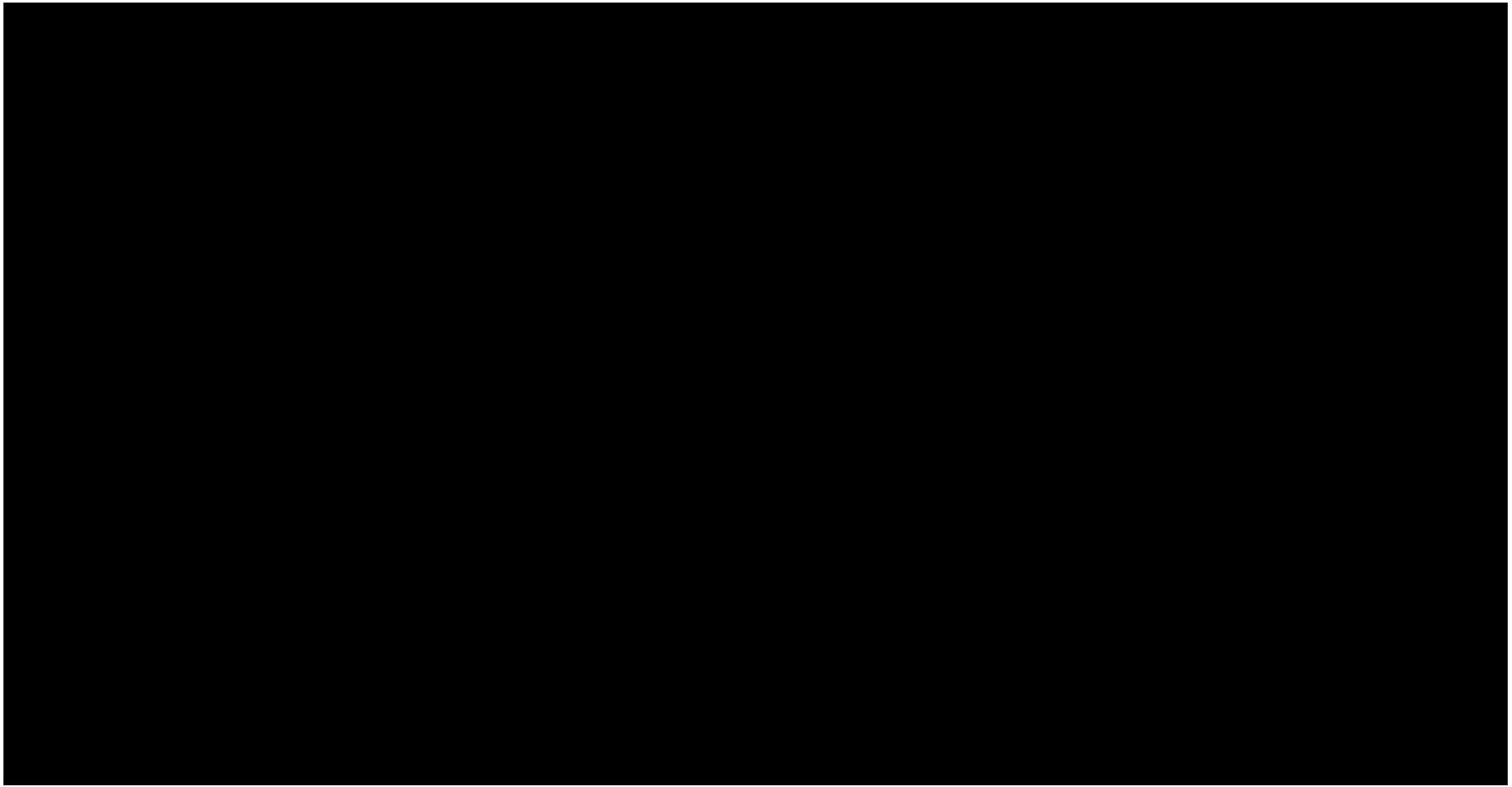
Metrics	YOLO	SSD	Classification
Recall	**0.9512**	0.7075	0.9355
Precision	0.806	0.6414	**0.9063**
mAP:@0.5	0.9533	**0.9535**	N/A
mAP:@0.95	**0.6439**	0.6412	N/A

Model Deployment

- Text-to-speech conversion with Pyttsx3
- Model deployment on Deepstack server

Demo

Experimental Setup



Some Roadblocks we Encountered:

- Dataset creation was hectic!
- Dataset Annotation was even more hectic!
- Currently trying to sort out dataset release...

Future Goals;

- Deployment in a mobile app
- Embedding in smart glasses

Thank You !

Code and config files are publicly available at: https://github.com/SteveKola/Sign-to-Speech-for-Sign-Language-Understanding

Preprint from ML4D@NeurIPS: https://arxiv.org/pdf/2111.00995.pdf