Singapore Sign (SgSL) Language Recognition through Computer Vision

Gloria Neo
15 Sept 2023

Contents

01

Problem Statement

Context

AUTOMATIC VOICE-TO-TEXT CAPTION GENERATION

FORBES > LEADERSHIP > DIVERSITY, EQUITY & INCLUSION

TikTok Makes Videos More Inclusive Of And Accessible To Deaf People With New Auto Captions Feature

Steven Aquino Contributor

Steven covers accessibility and assistive te

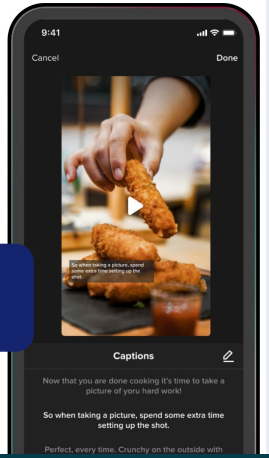
INCLUSIVITY AND ACCESSIBILITY

TikTok now supports auto captioning to bett

In a blog post published Tuesday, TikTok announced auto captions. The goal is to make the service more accessible to those in the Deaf and hard-of-hearing communities. "Inclusivity is important because when people feel included, they're more comfortable expressing themselves and engaging with their community. We're inclusive app environment, and that means tools that support our diverse community," of Creator Management and Operations at company's post.

With today's update, creators are able to tu editing page of the app, which then will aut spoken audio into text. Any generated capt the creator for better accuracy. On the view disable captions if they so choose.

TikTok says the captioning functionality is American English and Japanese, adding su coming months." Th ity organizations s is feature.



Captioning is making the world a more inclusive place

By **Matthew Johnston**

Published: January 27, 2020

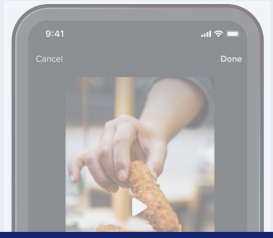
I only hear what I see

Last year, I became the **first deaf person to serve on an English jury**. When I received my jury summons, I never imagined I'd actually serve on a jury, let alone be selected as foreman. But after meeting with court officials, I convinced them that I could perform my civic duty through the use of subtitles. What they didn't realise was that others in the courtroom would benefit from the subtitles, too.

Context

FORBES > LEADERSHIP > DIVERSITY, EQUITY & INCLUSION

TikTok Makes Videos More Inclusive Of And Accessible To Deaf People With New Auto Captions Feature



The Hard-of-Hearing (HOH) community can now understand us.

But can we understand them?

AUTOMATIC VOICE-TO-TEXT CAPTION GENERATION

TikTok says the captioning functionality is
American English and Japanese, adding su
oming months." Th
ity organizations s
is feature.

more

Last year, I became the **first deaf person to serve on an English jury**. When I received my jury summons, I never imagined I'd actually serve on a jury, let alone be selected as foreman. But after meeting with court officials, I convinced them that I could perform my civic duty through the use of subtitles. What they didn't realise was that others in the courtroom would benefit from the subtitles, too.

Context

Technology and AI as enablers

For further transformation in the realm of media and communications

Holistic digital experience for the HOH community

Enable the hard-of-hearing (HOH) community to not just consume content, but to create content

Addressing the lack of resources for Singapore SL

Lack in existing technology or research tailored to SgSL

Problem Statement

- You have been engaged by Instagram in the development of a new feature enabling **automatic generation of text subtitles from live video signing**

- Use-case scenarios within Social Media and Communication Technology: Instagram reels functions, real-time captions during video calls, etc.

- To address this, your objective is to **explore a series of computer vision deep learning models that can perform recognition of Singapore Sign Language (SgSL)**

Project Scope

A multi-phased approach

PHASE 1

****CAPSTONE**

Alphabet Classification

- Identifying the 26 letters within the alphabet (Multiclass)

a) Image Classification

b) Object Detection on images and videos

PHASE 2

Word Detection

a) Build a dataset of SgSL word signing videos

b) Develop an object detection model for a corpus of single words

PHASE 3

Phrase, Sentence Detection

a) To study linguistic structure of SgSL

b) Use CV, NLP approaches to translate SgSL sequences into phrases and sentences.

02

Project Methodology

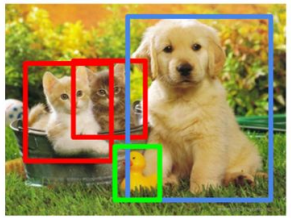
Image Classification VS Object Detection

Image Classification

CAT

predicting the class of one object in an image

Object Detection



CAT, DOG, DUCK

predicting the location of multiple objects in an image or video + classify them

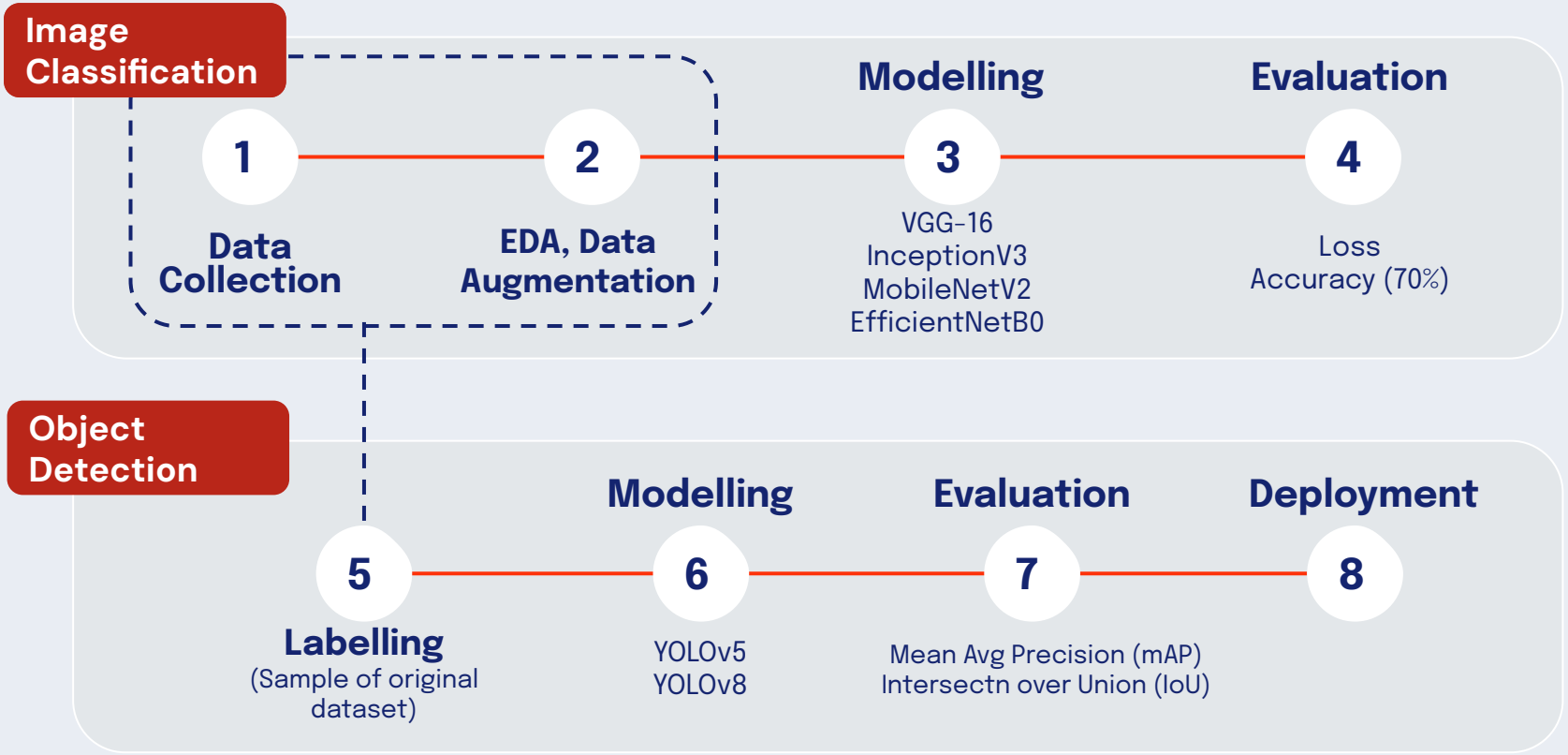
=

Object Localization

+

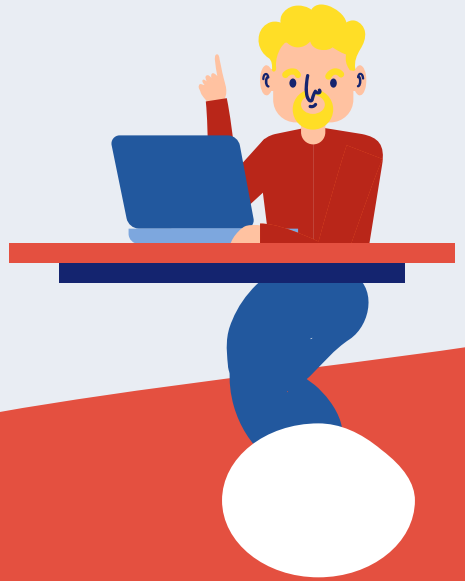
Object Classification

Project Methodology



03

Image Classification



1 Data Collection
2 EDA, Data Augmentation

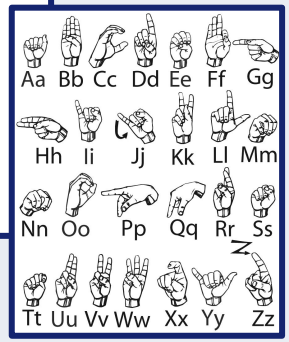
Image Classification

1 Data Collection → 2 EDA, Data Augmentation → 3 Modelling → 4 Evaluation

Data Collection

Datasets used:

- ASL Alphabet Dataset (Kaggle)
- ASL Alphabet Test (Kaggle)



EDA

Train Test Dataset Class Proportions

Preview of Dataset Image Selection

Brightness Distribution

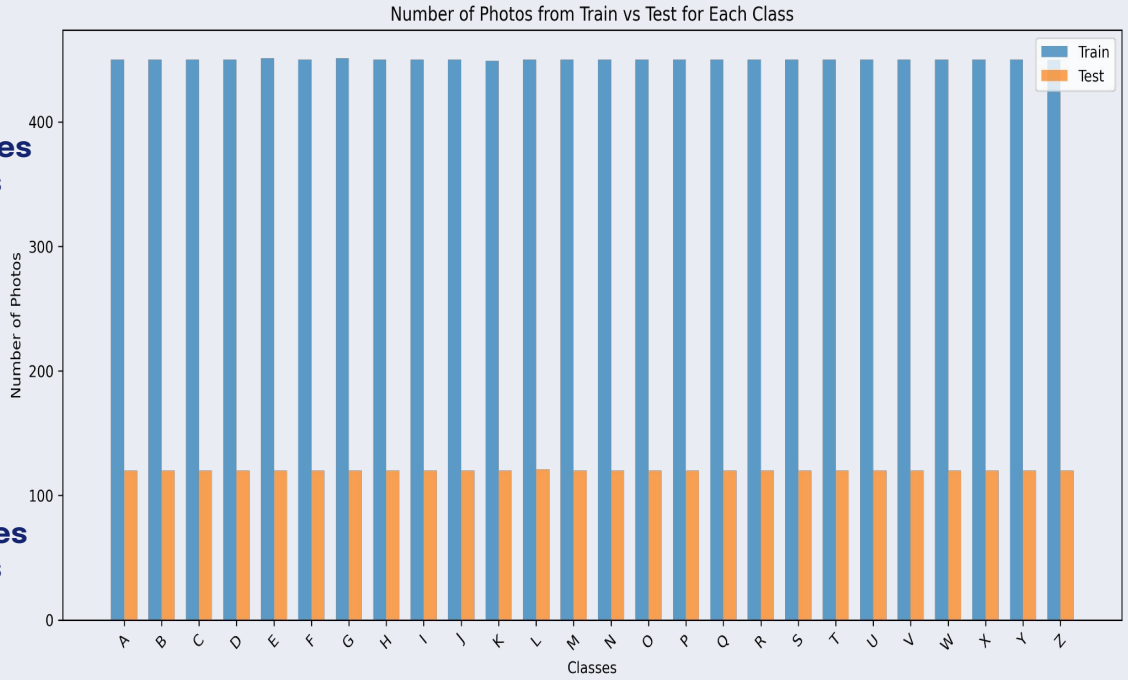
Data Augmentation

- Rotation, Shift, and Shear
- Zoom
- Horizontal and Vertical Flip
- Brightness Adjustment

Data Collection and EDA

Train vs Test Dataset

Train:
450 images per class

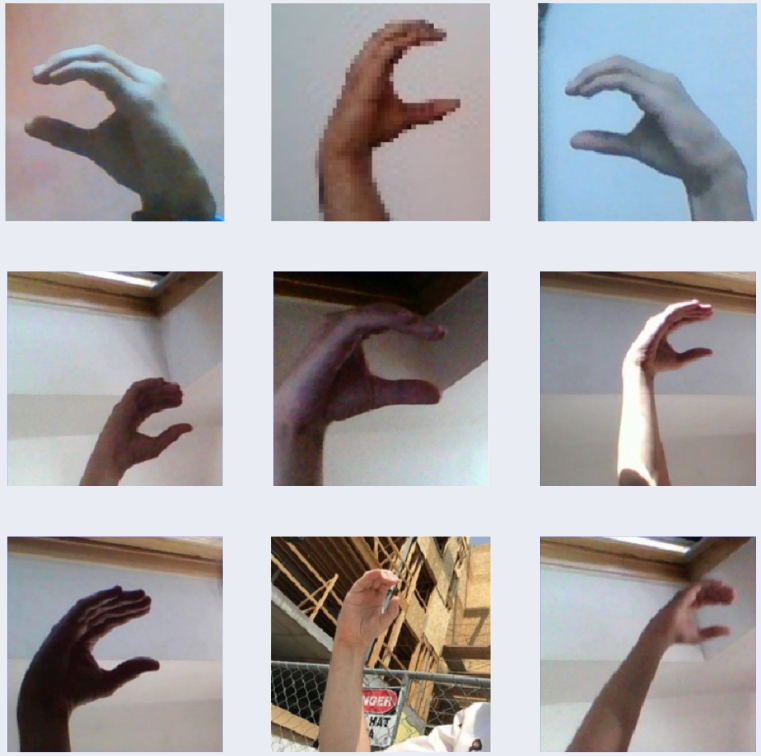


Test:
120 images per class

Multiclass Classification of 26 alphabet classes

Well-balanced classes → acceptable to use the metric of **accuracy** when evaluating our models

Preview of Dataset



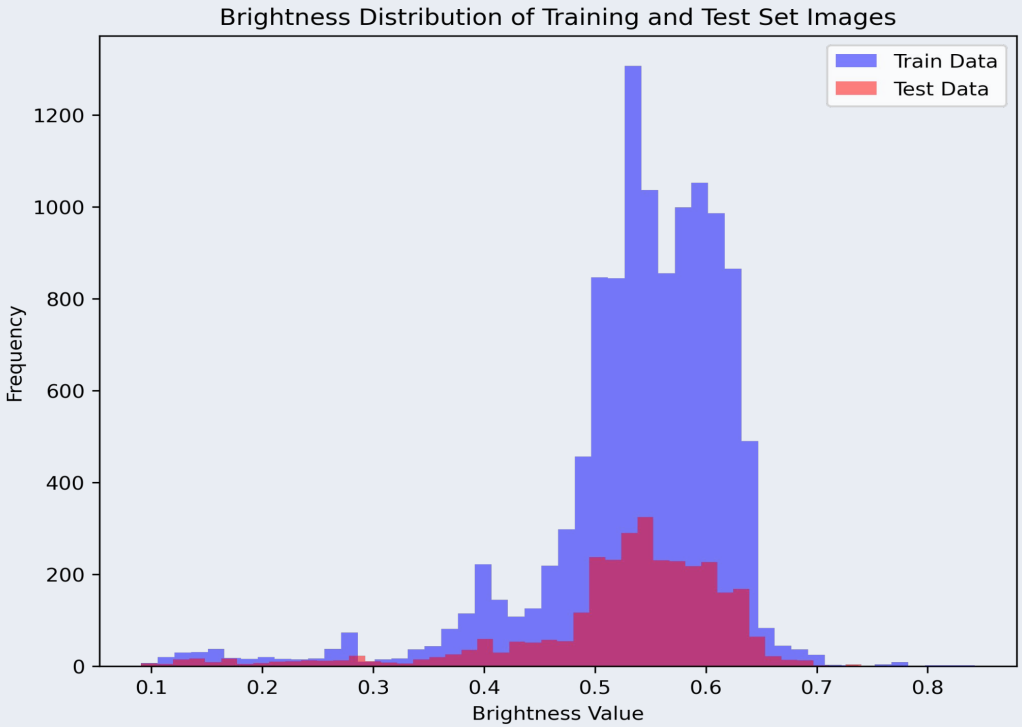
Preview of random selection from 'C' class folder for train dataset

Nuances captured within the dataset:

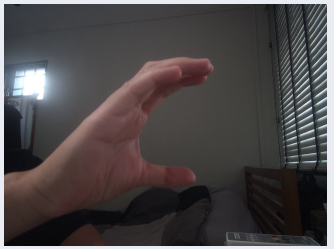
- In hand form or hand structure between different people
- In the way light falls on the hand

Brightness Distribution

Training and Test Set Images



- Left-skewed distribution
- Significant proportion of images with lower brightness value (< 0.5) → we only brighten images during data augmentation



Sample image (Train set)

Data Augmentation

Image Data Generator was used to perform the following techniques to our Training dataset:

- Rotation, Shift, and Shear
- Zoom
- Horizontal and Vertical Flip
- Brightness Adjustment

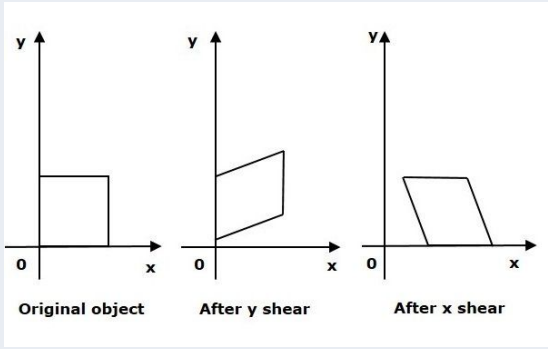


Diagram to explain Shearing

3 Modelling

Image Classification

1. Data Collection
2. EDA, Data Augmentation
3. Modelling
4. Evaluation

Model Input

Data Augmented Images

Experimentation on Base Model

Pre-trained Models

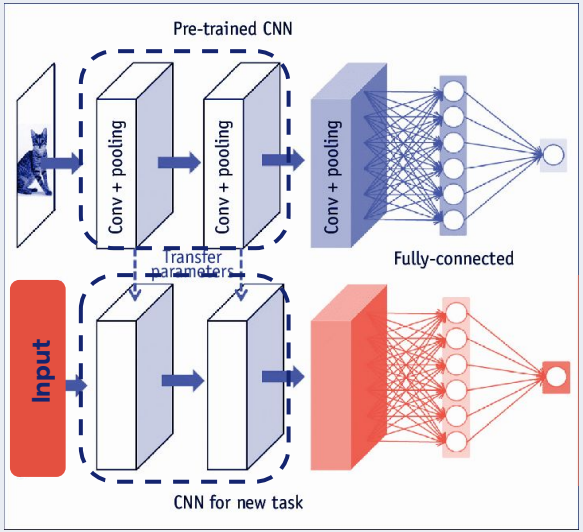
- VGG -16
- Inception V3
- Mobile NetV2
- Efficient NetB0

Custom Fully Connected Layers

- Additional hidden layer with 64 nodes
- Output layer for 26 classes
- Softmax Activation Function

Selected Parameters

BATCH_SIZE = 64
IMG_SIZE = 224
EPOCHS = 50
DROPOUT_RATE = 0.2
LEARNING_RATE = 0.001

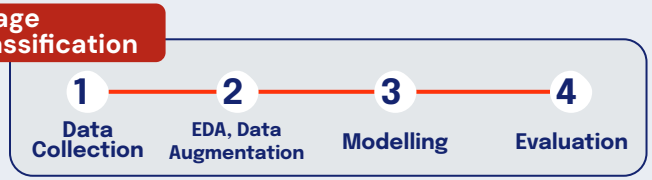


Transfer Learning

4 Model Evaluation

Metrics

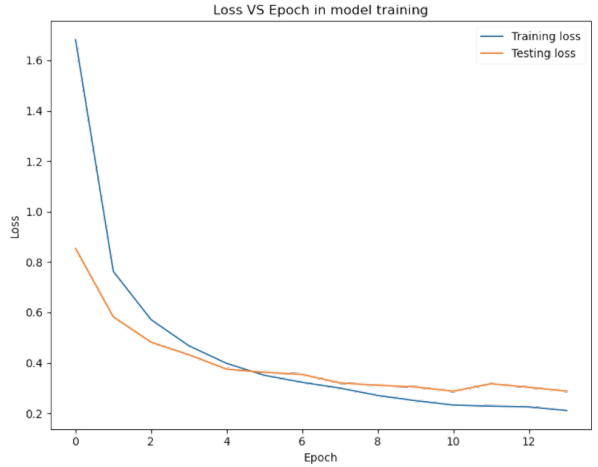
- Accuracy (70%)
- Loss

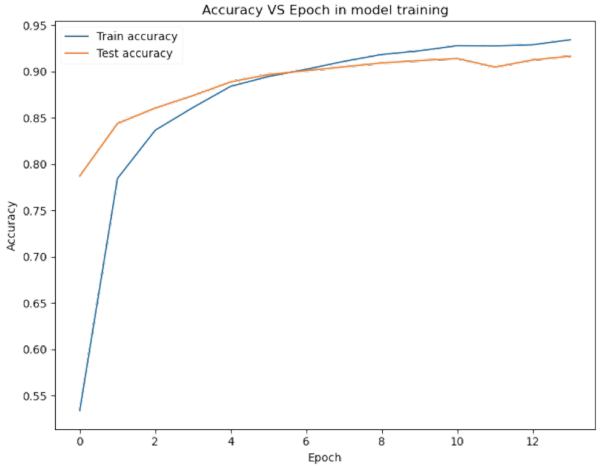


No.	Model Used	Loss	Accuracy	Conclusion
1	**VGG-16**	0.4052 (Train) 0.4675 (Test)	0.8761 (Train) 0.8718 (Test)	No over or underfitting
2	**InceptionV3**	0.4607 (Train) 0.4606 (Test)	0.8483 (Train) 0.8597 (Test)	Slight underfitting
3	**MobileNetV2**	0.2327 (Train) 0.2874 (Test)	0.9280 (Train) 0.9141 (Test)	Slight overfitting
4	**EfficientNetB0**	0.1263 (Train) 0.2434 (Test)	**0.9612 (Train)** **0.9314 (Test)**	Slight overfitting Best score

MobileNetV2

Best Epoch: 11
Train Loss: 0.2327
Validation Loss: 0.2874
Train Accuracy: 0.9280
Validation Accuracy: 0.9141

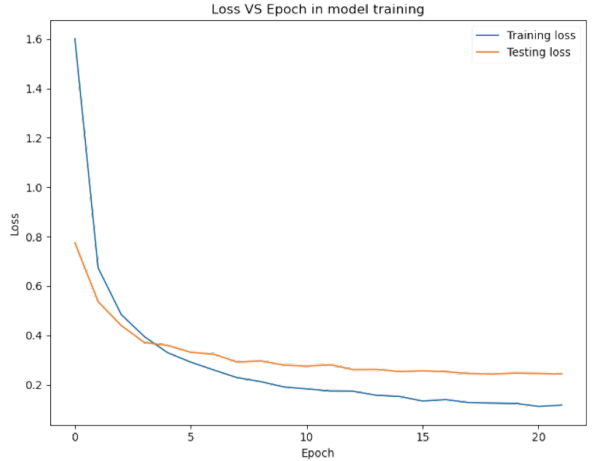


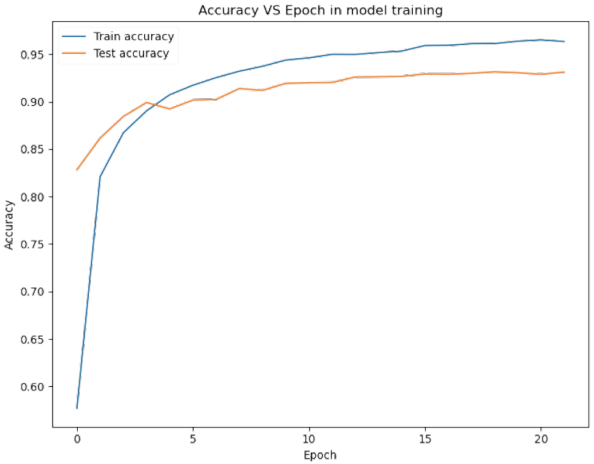


EfficientNetB0

Best Epoch: 19
Train Loss: 0.1263
Validation Loss: 0.2434
Train Accuracy: 0.9612
Validation Accuracy: 0.9314

- Avg training time per epoch
  EfficientNetB0: 580s
  EfficientNetB4: 1800s
  EfficientNetB7: 3000s



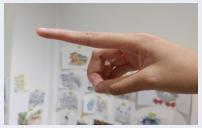


Model Inference

Model inference with best EfficientNetB0 model

Top 1 Prediction:
C, Probability: 0.7620

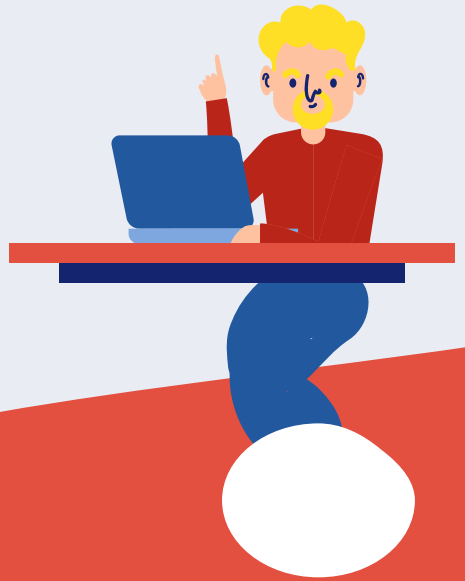
Top 2 Prediction:
P, Probability: 0.1418



Top 3 Prediction:
Q, Probability: 0.0736

04

Object Detection



Object Detection Workflow

Labelling

Smaller sample (262 images) selected from full dataset for labelling via Roboflow

Pre-trained Models

YOLOv5s	**YOLOv8s**
IMGSZ = 460 EPOCHS = 50	**IMGSZ = 640** **EPOCHS = 50**

Experiment with Single-shot object detection architectures

Model Evaluation

Mean Average Precision (mAP)

Precision

Recall

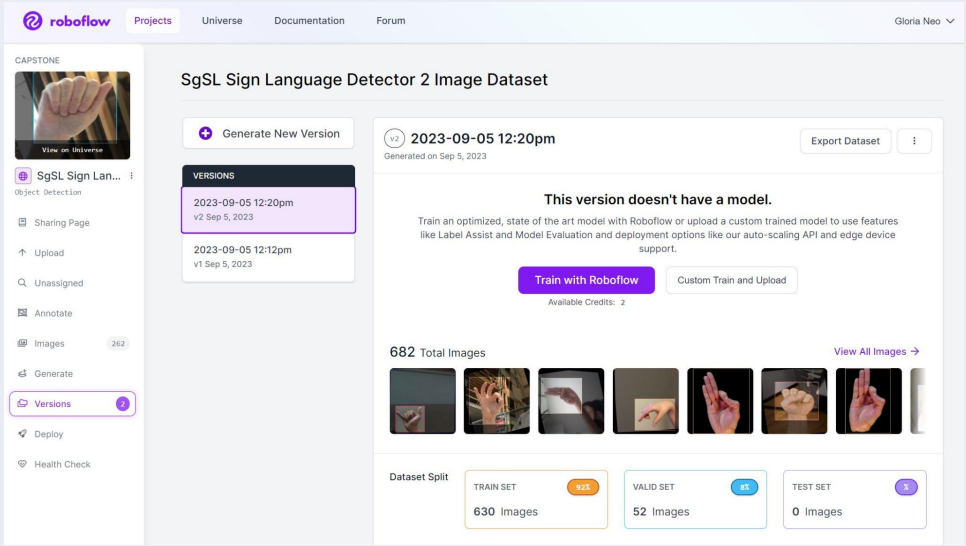
Deployment

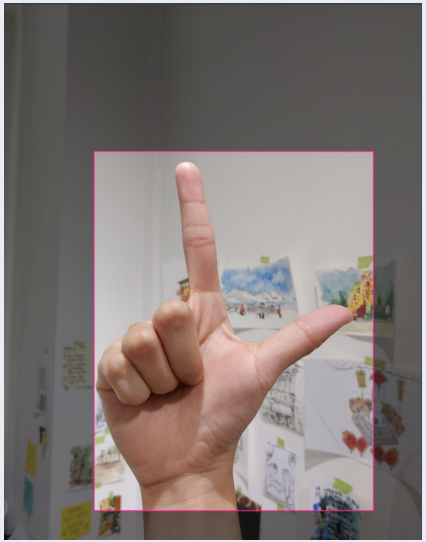
Deployment of YOLOv8 model on Streamlit

-Images
-Videos

5 Data -Labelling

- Labelling is done on Roboflow platform
- Data sample: **262 images** (pre-augmentation)
- After data augmentation: Train: 630 images, Test: 52 images



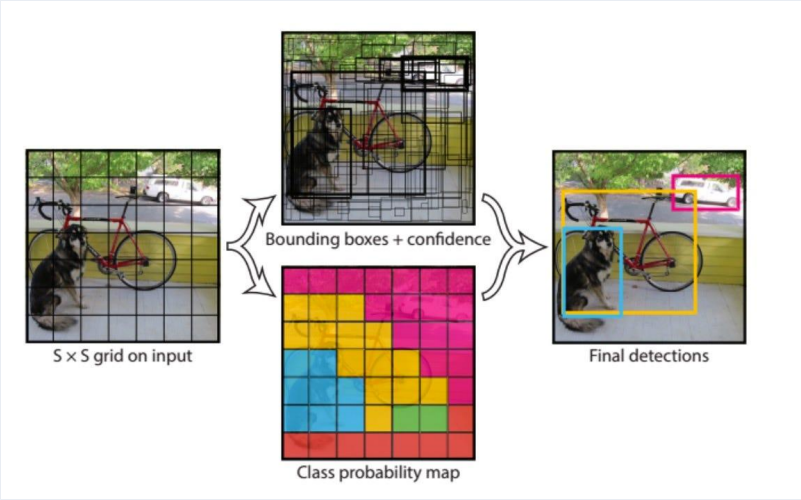


6 Model - YOLO

Object Detection

5 Labelling — 6 Modelling — 7 Evaluation — 8 Deployment

- YOLO - You Only Look Once
- Single-shot object detection - make predictions about the presence and location of objects in the image in a single pass
- Grid Division > Bounding box prediction > Class prediction > Thresholding > Non-max Suppression > Output



7 Metrics

- **Mean Average Precision (mAP)**
  - mAP is calculated by taking the mean AP over all classes and/or overall IoU thresholds
  - Used to assess model's overall performance across all classes

- **Precision**, **Recall**
  Precision: Proportion of true positives among all positive predictions made by the model

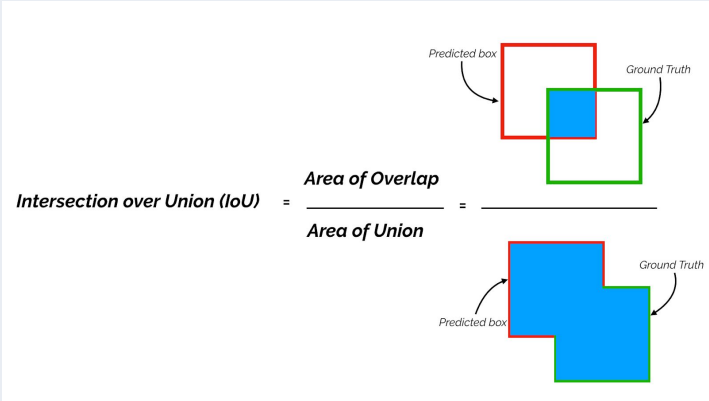
  Recall: Proportion of true positives among all actual positive instances in the dataset.

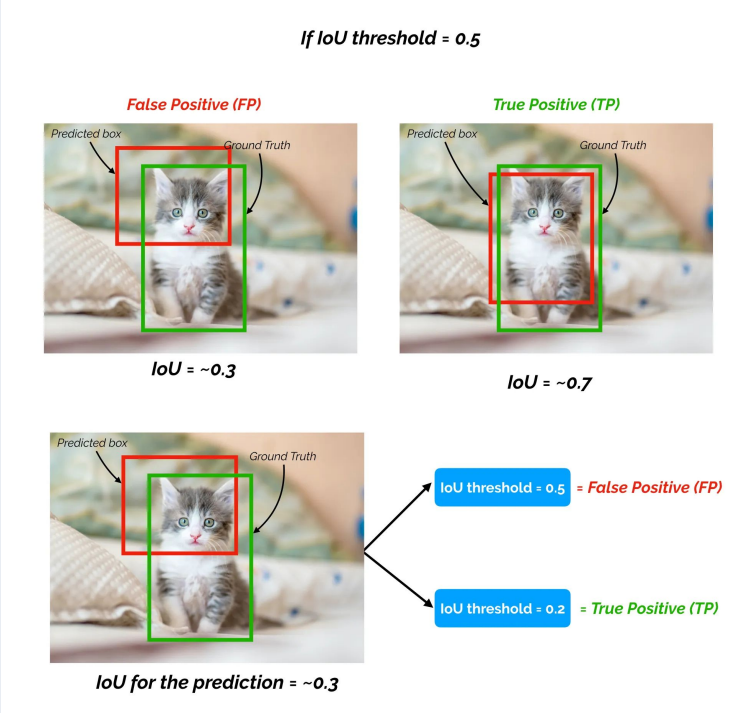
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Intersection over Union

- IoU value is used to measure the overlap between the predicted bounding box and the ground truth bounding box
- Used as a threshold to determine whether a detection is correct or not.





7 Model Evaluation

Object Detection

5 Labelling — 6 Modelling — 7 Evaluation — 8 Deployment

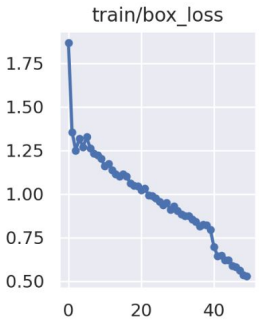
Model Used	Precision	Recall	mAP at IoU 0.5
YOLOv5 (YOLOv5s - small) (different dataset) IMGSZ = 460 EPOCHS = 50	0.248 (Train) 0.333 (Test)	0.462 (Train) 0.314 (Test)	0.367 (Train) 0.274 (Train)
YOLOv8 **(YOLOv8s - small)** **IMGSZ = 640** **EPOCHS = 50**	**0.896 (Train)** **0.896 (Test)**	**0.655 (Train)** **0.656 (Test)**	**0.848 (Train)** **0.848 (Train)**

Model Results - YOLOv8

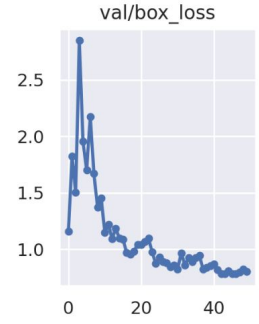
Loss

MSE

Train

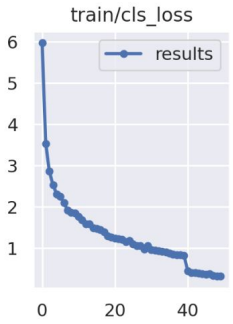


Test

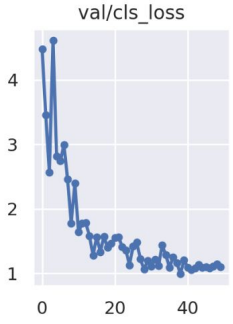


Cross-entropy loss

Train

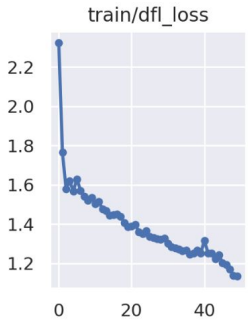


Test

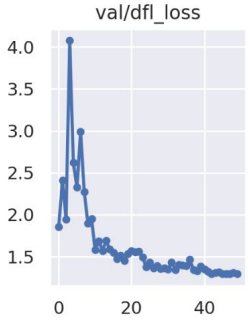


Distribution Focal Loss

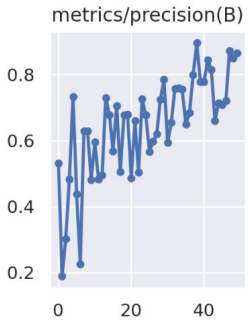
Train

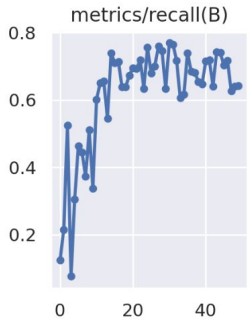


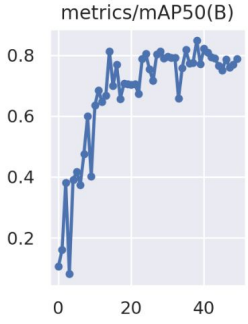
Test

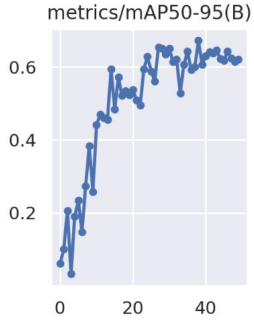


Metrics - Best



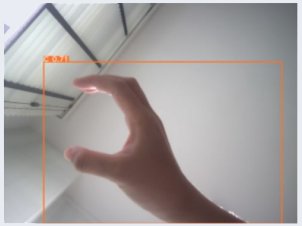




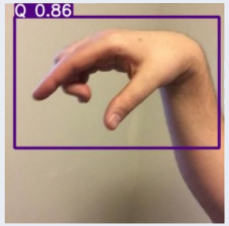


Model Inference - YOLOv8s

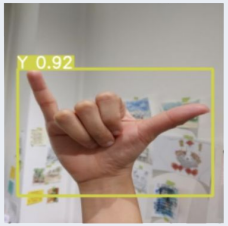
Actual Class: C (2).jpg



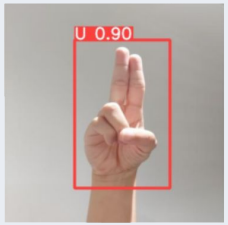
Actual Class: Q.jpg



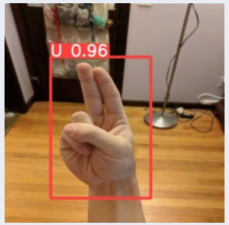
Actual Class: Y.jpg



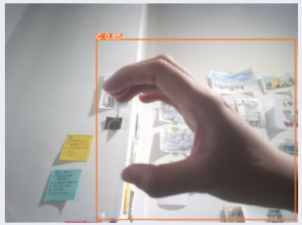
Actual Class: U.jpg



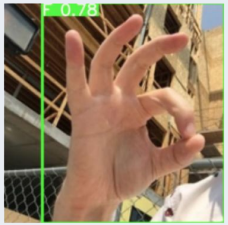
Actual Class: U (1).jpg



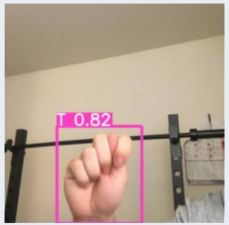
Actual Class: C (1).jpg



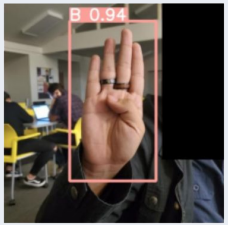
Actual Class: F.jpg



Actual Class: T.jpg



Actual Class: B.jpg



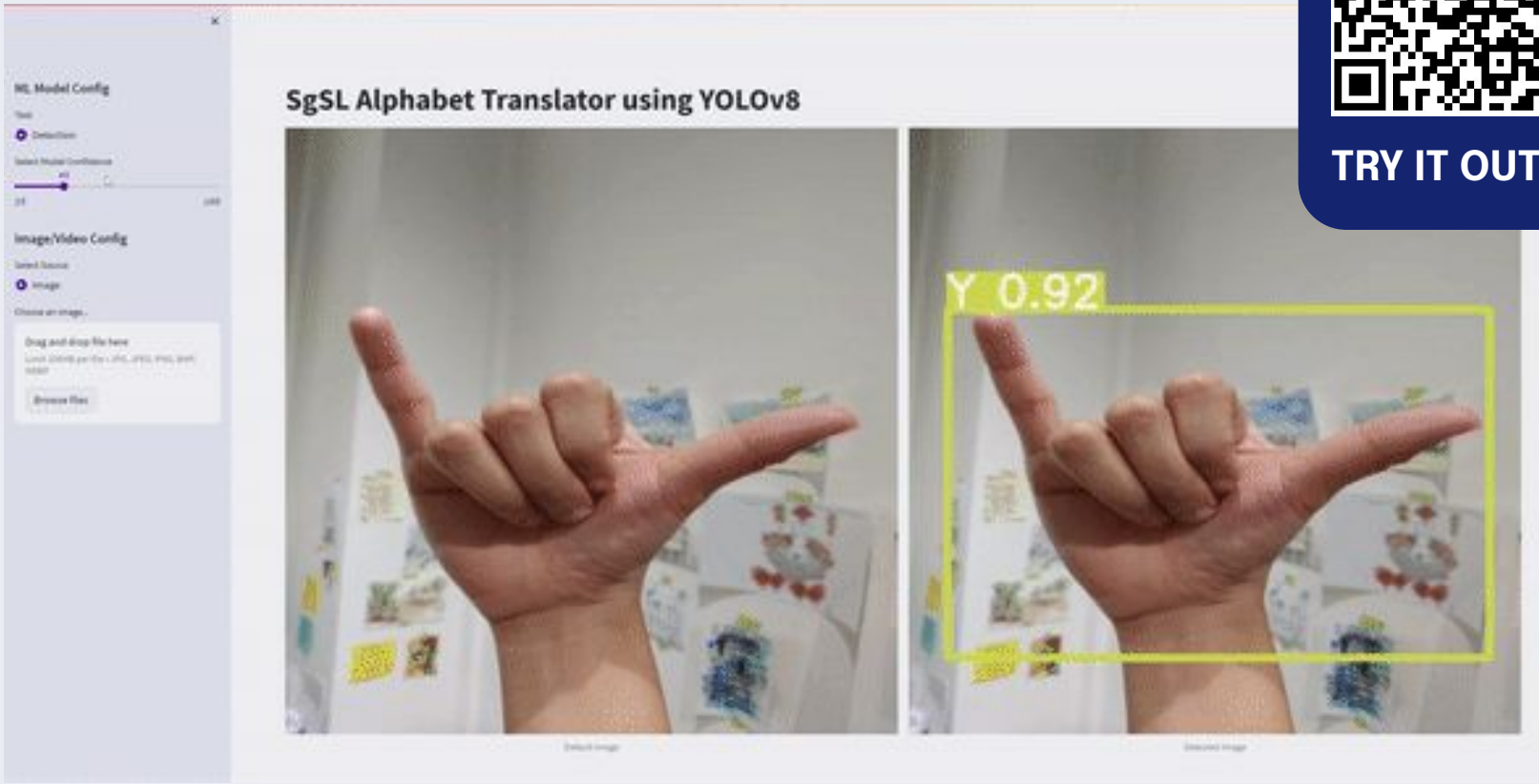
Model Inference - YOLOv8s

05

Streamlit Deployment

Streamlit Deployment

https://sgsl-recognition.streamlit.app/



TRY IT OUT!

06

Conclusion, Future Works

Conclusion

Proof of Concept (POC)

This capstone project demonstrated the potentials that computer vision in sign language recognition could bring

Decent Model Performance

Image Classification:
Best model scored >90% for both train and test accuracy

Object Detection:
mAP at IoU=0.5: 84%

Limitations and Recommendations

Data Availability

Lack of availability of SgSL sign language datasets

Important for capturing nuances in local handsigning

Computational Power, Time Constraints

For object detection and future approaches:

Other versions of YOLO models should be explored

Ambiguity and Context

Contextual cues and facial expressions introduce ambiguity

To explore models that recognize nuanced meanings, SgSL linguistic structure

THANK YOU!

Email:
gloria.nxe@gmail.com

Linkedin:
https://www.linkedin.com/in/gloria-neo/

Github:
https://github.com/GloriaNeo