

# HAND SIGN LANGUAGE RECOGNITION WITH ARTIFICIAL INTELLIGENCE

Using “You Only Look Once” (Yolo) model as a case

Bui, Hien Minh

Bachelor Thesis  
Degree Programme in Business Information Technology  
Bachelor of Business Administration

2022

Business Information Technology  
Bachelor of Business Administration

---

<b>Author</b>	Hien Minh Bui	<b>Year</b>	2022
<b>Supervisor</b>	Pekka Reijonen		
<b>Title of Thesis</b>	Hand sign language recognition with artificial intelligence		
<b>Number of pages</b>	29		

---

This thesis introduces information on sign language and the culture around it. Additionally, what was still lacking in the other existing research before and after Yolo is introduced including my unique contributions to this topic using the constructive research method.

In this thesis, the information about sign language and deaf-and-mute community was referenced from related books and research to provide the readers reliable insight about this new field. Furthermore, the technical knowledge was collected from Internet sources and scientific articles to provide an up-to-date and accurate information at the time this thesis was written. The data source that was used to train the Yolo model was manually collected from the YouTube videos to ensure its diversity.

For demonstration, a website application was created to evaluate the reliability of using the Yolo model for translating American sign language alphabets. Finally, the conclusion of this thesis work indicate that Yolo version 3 was not an optimal solution for sign language despite its outstanding speeds and efficiency in static object detection.

Key words

Sign language, Deep Learning, Yolo, Objects detection

# CONTENTS

Abstract

## SYMBOLS AND ABBREVIATIONS

1. INTRODUCTION.....	6
1.1 Deaf-and-Mute Community and Sign Language.....	6
1.2 Is Hearing Aid a Solution? .....	6
2. MOTIVATION, RESEARCH APPROACH, AND UNIQUE CONTRIBUTION.....	8
2.1 Motivation .....	8
2.2 Research Approach .....	9
2.3 Unique Contribution .....	9
3. THEORY AND FUNDAMENTAL INFORMATION.....	11
3.1 Introduction To Artificial Intelligence and Machine Learning .....	11
3.2 Introduction To Yolo Algorithm.....	11
3.2.1 History of Development.....	11
3.2.2 What Makes Yolo Stand Out? .....	12
3.2.3 How Is Yolov3 Different?.....	12
3.3 Pros and Cons of The Approach.....	14
4. CONSTRUCTION OF THE YOLO MODEL.....	16
4.1 Dataset .....	16
4.1.1 Data Source .....	16
4.1.2 How Does Yolo Work with This Dataset?.....	18
4.1.3 Labeling Tool.....	19
4.2 Training and Evaluation .....	20
4.2.1 Training .....	20
4.2.2 Evaluation .....	21

5. CONCLUSION .....26

BIBLIOGRAPHY .....27

## SYMBOLS AND ABBREVIATIONS

ASL	American Sign Language
CRA	Constructive Research Approach
ML	Machine Learning
AI	Artificial Intelligence
Yolo	You Only Look Once
GPU	Graphics Processing Unit
CPU	Central Processing Unit

## 1. INTRODUCTION

In this chapter, I give you a brief information about the deaf-and mute community, the culture around sign language, and the current situation with the hearing aid.

### 1.1 Deaf-and-Mute Community and Sign Language

Deaf and mute is usually considered as a disability by the majority of normal speaking people. In fact, the deaf people do not consider themselves disabled, deafness just mean that they have different way of communication. The awareness of deaf-and-mute community is rising lately, the world has an anniversary month to pay tribute to the deaf. (Guth 2020.)

The sign language, especially American Sign Language (ASL) has all the fundamental linguistic components that form a normal spoken language, such as phonetics, phonology, syntax, semantics and pragmatics. The sign language also has its own grammar which is different from the oral language of the local area where it is commonly used. (Pribanić 2006, 16.) For example, the words in the question “what is your name?” can be ordered differently between in ASL and BSL (British sign language). It made the diversity of the sign culture, there are approximately 300 different sign languages worldwide today, which may surprise many people who understand that there is only one sign language used internationally (Sign Solutions 2021). Despite that, all sign languages, in general, share common communicating manners, which are facial expression, hand postures, hand gestures, and a series of postures (Pribanić 2006, 10-12).

### 1.2 Is Hearing Aid a Solution?

According to research executed by World Health Organization in 2021, there would be one out of ten people having problem with hearing loss by 2050 whereas only hard-of-hearing people can benefit from hearing assistive device and the deaf people need to use sign language for communicating (World Health Organization 2021). Deaf-and-mute community is familiar with hearing aid, a device which helps the hard-of-hearing people to hear by amplifying the noise. Despite of the benefit brought by

the hearing aid, many deaf and hear-of-hearing people refuse to use it. The reasons for this issue are lack of robustness and consistency of hearing aid, further, there are no standard methods or tools to assess the quality of this device. In addition to the stability of the device, many people fitted with a hearing aid decide not to keep using it because they found it difficult to insert the ear mould and cope with signals in noise. (McCormack & Fortnum 2013, 360-361.)

Even though, many improvements have been made for the hearing aid, such as increased comforts in wearing experience, reduction in digital noise, enhancement for digital speech, smaller size, etc. there is still evidence of underuse of hearing aid among the deaf-and-mute community (McCormack & Fortnum 2013, 361). This situation may lead to more severe problems, according to McCormack and Fortnum (2013, 361), Kochkin (2012) found evidence showing that using hearing aid improves the relationships among family members, stability in emotions, the feeling of having the control of the life events, etc., which mean that if a person who has hearing difficulty but refuses to use the hearing aid, there is a higher chance of causing depression and anxiety symptoms (McCormack & Fortnum 2013, 361; Gopinath et al. 2009, 1306-1308).

## 2. MOTIVATION, RESEARCH APPROACH, AND UNIQUE CONTRIBUTION

This chapter gives you a clearer view on my motivation for the topic and existing research on the same topic but having different approaches. Besides, I would like to walk you through my research method as well as my unique contribution.

### 2.1 Motivation

There are many scholars who researched the similar topics before, however, based on my experience in reading those research papers, I categorized the previous research into two types, those that were completed before and after the You Only Look Once (Yolo) algorithm was introduced. Below are two examples of each type to provide clearer view on what are still lacking from the previously completed research.

The earlier research, recognizing hand sign using other techniques than Yolo model, makes the final products become too slow to translate the language. For example, according to Bhowmick and Kumar (2015, 406) from Tezpur University, they used the Hidden Markov model to translate sequence of posture/gestures with multiple middle process such as gesture module and image acquisition, tracking, and classification (Bhowmick & Kumar 2015, 406). As you can see, this approach involved a lot of computing process which certainly affect the speed of the performance of the products which use this research as its core technology.

The latter research, recognizing hand sign using the Yolo model, did not have the application product to demonstrate the result nor the evaluation of various shapes of hands. For example, according to Nanda (2020, 44-47) from Purdue University, he used Yolov3 for recognizing ASL with his own hands the data of which is also used to train the Yolo model, which would probably not work with other different hands shapes. In that document, he also did not build an end-to-end product to demonstrate the model performance, but just testing the performance with his own photo with different backgrounds (Nanda 2020, 44-47). Therefore, the research still did not manage to evaluate the performance of the Yolo model in terms of accuracy on unseen data, neither did it demonstrate the applicability of the Yolo model.



Therefore, I wanted to do this project using the Constructive Research Approach (CRA), not only for myself to discover a new domain knowledge and raise the awareness about the deaf-and-mute community, but also to humbly contribute to the prior scientific research in an effort to build two ways communication between two communities having different communication manners.

## 2.2 Research Approach

In this thesis document, I used the CRA to demonstrate and verify the helpfulness of modern technology, especially Machine Learning (ML) and Artificial Intelligence (AI), in an effort to translate the sign language by building an application to capture a person photo signing ASL alphabet and detect the letter.

But what is CRA? CRA is primarily developed for the field of Business Administration, it can potentially be applied to other fields as well. This research method gained the more and more attention from, not only business experts but also, engineering experts, especially in the information system field. (Lukka 2003, 83.) What makes CRA match this thesis's goal? One of purposes of CRA is to exemplify the usability of a new method by the doing market test on an innovative product or application (Rautiainen, Sippola & Mättö 2017, 1-3). The CRA requires the learner to create a module, tool, or method, which must have a standard of maturity beyond a simple case study. The researcher needs to come up with a solution, the maturity of which could be varied, for transdisciplinary problem to prove the applicability of a theory or demonstrate possibility for collaboration between several industrial or scientific fields. (McGregor 2018, 7-8.) In this case, tackling the topic of this thesis requires a wide range of knowledge of sign language, society, science, and technology (the core of this thesis). Based on the above features of CRA and my current studying purpose toward the topic data science, I find CRA the most suitable research method for this thesis.

## 2.3 Unique Contribution

In this thesis, I had three unique contributions compared to previous similar topic research. The first one is the dataset, which is collected from more than 50 YouTube

tutorial videos for sign language, uniquely created by me with data augmentation technique to deal with as many hands shapes and lightning conditions as possible from the unseen data for testing purpose.

Secondly, the web application, which is accessible by internet, is built with Yolo model as the core technology to demonstrate the applicability of the model and evaluate the performance of the model with real volunteering people.

Lastly, my model evaluation method used the real experience of the volunteers who used the mentioned web application and gave their feedbacks, this way which allowed to assess the model performance on the unseen data with various hands shape and lightning condition.

### 3. THEORY AND FUNDAMENTAL INFORMATION

In this chapter, I discussed general definition of AI and ML as well as the prerequisite knowledge of Yolo model.

#### 3.1 Introduction To Artificial Intelligence and Machine Learning

AI is a broad field which focuses on training the computer to do certain tasks with as high an accuracy as, or even higher than, human beings. ML is a subfield of AI, ML's purpose is to give the computers to learn without being explicitly programmed. (Brown 2021.) This section, I introduced how to use ML to train a "model" that can recognize the ASL alphabets hand postures.

#### 3.2 Introduction To Yolo Algorithm

In this subchapter, you will get the basic information about Yolo algorithm such as its development history, what made Yolo different from the previous approach and how Yolo third version differed from its "ancestor".

##### 3.2.1 History of Development

First, let's get some insight into the developing history of Yolo model. It is a ML model built for object detection purpose. Joseph Redmon at University of Washington was the pioneer in this project, he was involved in the first three versions of Yolo project, before stopping contribution to the development of Yolo because of the "military application and privacy concern". (Yuan 2020.)

In 2020, Alexey Bochkovski continued the research and publish the Yolo version four; in the same year, version five was developed, but there were many controversies involved with that version. In this thesis, I used the Yolov3 to train the ML model to detect the hand sign language as a way to show respect to the first author of this project.

### 3.2.2 What Makes Yolo Stand Out?

When the first Yolo version was published in 2016, it outperformed all the prior object detection methods. In the first published research, the team of Joseph Redmon stated three reasons that make Yolo algorithm more optimal than the previous methods. (Redmon, Divvala, Girshick & Farhadi 2016, 1-2.)

Firstly, Yolo is fast because they see the object detection as a regression model and remove the complicated pipeline. Secondly, Yolo looks at an image as a whole so that it can encode the contextual information about the object appearing in the image, while the Fast Region Convolutional Neural Network (Fast R-CNN) cannot see the larger context of an image because this method dismissed the context in the background. Thirdly, Yolo has its own mechanism to generalize the training images by automatically adding large margin to the training images, hence, it can handle the unexpected inputs from the testing data very well. (Redmon, Divvala, Girshick & Farhadi 2016, 1-2.)

### 3.2.3 How Is Yolov3 Different?

The first thing noticed is that the network of darknet-19 in Yolov2 is now called darknet-53. The reason for this is that in this version the architecture of Yolov3 has 53 convolutional layers (See the 2 images below) It is a bigger architecture, hence, improve the outstanding accuracy of its ancestor, but this version is still fast enough to outperform the other methods, such as ResNet-101 and ResNet-152. (Redmon & Farhadi 2016, 1-7.)

Type	Filters	Size/Stride	Output
Convolutional	32	3 × 3	224 × 224
Maxpool		2 × 2/2	112 × 112
Convolutional	64	3 × 3	112 × 112
Maxpool		2 × 2/2	56 × 56
Convolutional	128	3 × 3	56 × 56
Convolutional	64	1 × 1	56 × 56
Convolutional	128	3 × 3	56 × 56
Maxpool		2 × 2/2	28 × 28
Convolutional	256	3 × 3	28 × 28
Convolutional	128	1 × 1	28 × 28
Convolutional	256	3 × 3	28 × 28
Maxpool		2 × 2/2	14 × 14
Convolutional	512	3 × 3	14 × 14
Convolutional	256	1 × 1	14 × 14
Convolutional	512	3 × 3	14 × 14
Convolutional	256	1 × 1	14 × 14
Convolutional	512	3 × 3	14 × 14
Maxpool		2 × 2/2	7 × 7
Convolutional	1024	3 × 3	7 × 7
Convolutional	512	1 × 1	7 × 7
Convolutional	1024	3 × 3	7 × 7
Convolutional	512	1 × 1	7 × 7
Convolutional	1024	3 × 3	7 × 7
Convolutional	1000	1 × 1	7 × 7
Avgpool		Global	1000
Softmax			

Figure 1. Yolov2 (Redmon &amp; Farhadi 2016, 6)

Type	Filters	Size	Output
Convolutional	32	3 × 3	256 × 256
Convolutional	64	3 × 3 / 2	128 × 128
Convolutional	32	1 × 1	
Convolutional	64	3 × 3	
Residual			128 × 128
Convolutional	128	3 × 3 / 2	64 × 64
Convolutional	64	1 × 1	
Convolutional	128	3 × 3	
Residual			64 × 64
Convolutional	256	3 × 3 / 2	32 × 32
Convolutional	128	1 × 1	
Convolutional	256	3 × 3	
Residual			32 × 32
Convolutional	512	3 × 3 / 2	16 × 16
Convolutional	256	1 × 1	
Convolutional	512	3 × 3	
Residual			16 × 16
Convolutional	1024	3 × 3 / 2	8 × 8
Convolutional	512	1 × 1	
Convolutional	1024	3 × 3	
Residual			8 × 8
Avgpool		Global	
Connected		1000	
Softmax			

Figure 2. Yolov3. (Redmon &amp; Farhadi 2018, 2)

The next thing noticed is the way Yolov3 predict the class (1) is different from Yolov2. In Yolov2, it used softmax as the activation function (see the below for softmax formula), to generate a vector of probabilities value for each bounding box, then choose the class that have the highest probability for each bounding box (Redmon & Farhadi 2018, 1-5; Radevic 2020). This method assumes that each bounding should have exactly one class with it, which is not always the case in the real-life data. Yolov3 helps to overcome the obstacle. (Redmon & Farhadi 2018, 1-5.)

$$P = \langle p_1 \quad \dots \quad p_N \rangle = \frac{\langle e^{z_1} \quad \dots \quad e^{z_N} \rangle}{\sum_{j=1}^N e^{z_j}}$$

*Softmax activation function, With “z” is the output value of the convolutional neural, “N” is the number classes, “p” is*

*the probability of each class, and “P” is the vector of probability values (Radecic 2020).*

Instead, Yolov3 used the logistic regression as activation function for each class independently in each bounding box, then apply binary cross-entropy loss function to evaluate the prediction accuracy. This way allows Yolov3 model to better the dataset having overlapping labels. (Redmon & Farhadi 2018, 1-5.)

$$\hat{Y} = \delta(z) = \frac{1}{1 + e^{-z}}$$

*Logistic Regression formula, also known as sigmoid function. With “z” is the output value of the convolutional neural network (Loiseau 2020).*

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^N Y_i \log(\hat{Y}_i) + (1 - Y_i) \log(1 - \hat{Y}_i)$$

*Binary Cross-entropy loss function. With “ $\hat{Y}$ ” is the predicted label for from the Logistic Regression and “Y” is the true label (Loiseau 2020).*

As mentioned in the section 1.1, hand postures are seen as static images, unlike a gesture-formed words in sign language, which is created by series of postures, which are not able to be detected by Yolo model.

(1) Each class in Yolo model is the output predicted or labeled object.

### 3.3 Pros and Cons of The Approach

About the pros of using Yolo, firstly, as mentioned in the previous sections, Yolo is more accurate compared to the previous object detection methods. Secondly, Yolov3-tiny version model can work in real-time video without Graphics Processing Unit (GPU) supports, video can be up to 220 FPS with GPU supports. (Redmon & Farhadi 2018, 1-5.) Finally, it can detect up to 9000 classes, this number is way over

the decent amount of vocabulary needed for daily conversation (Redmon & Farhadi 2016, 1-7).

About the cons of using Yolo, firstly, Yolo takes color into account, so the color of skin may affect the result. Secondly, in terms of speed-wise for standard YoloV3 version, Central Processing Unit (CPU) core can take 6-12 seconds for each image (Redmon & Farhadi 2018, 1-5). Finally, the most drawback of Yolo algorithm, Yolo only works well for posture-based words, not with the dynamic gesture-based words. Therefore, for demonstration purpose in this thesis, I took the final frames of a dynamic gesture alphabets to represent them. (Nanda 2020, 12.)

## 4. CONSTRUCTION OF THE YOLO MODEL

This is chapter 4, the main chapter of the thesis, where I tell you more about my construction process. Two main parts are my dataset and how I used it to train and evaluate the model.

### 4.1 Dataset

In this subchapter, you will know the source of my training data and how Yolo model adapt to my specific dataset. In addition to those, I also introduced you the labelling tool I used to label my dataset.

#### 4.1.1 Data Source

The data used in this project is collected from Youtube ASL alphabet tutorial. I captured the screenshot of each alphabet in the videos. I collected data from more than 50 different tutorial videos, each video has at least 1 image for each alphabet. See examples below.



Figure 3. Data sample collected from YouTube video, from left to right, reference (Dellis 2020; ASL That 2013; Learn How to Sign 2020)

However, Yolo demand more than that number of images for each class to train effectively, there should be at least approximately 1000 for each class for Yolo to train optimally (Warden 2017). Then, I implemented the data augmentation (2) for each image, the technique I used is to change the contrast, blurriness, brightness, flipping, etc (Shorten & Khoshgoftaar 2019, 7). After augmenting the collected image, I had a new dataset that has more than 1300 images for each class (alphabet). Each original image has its 23 other variances, see the below image as



an example. This method allows to train the model better with different camera quality and unexpected natural or digital lightness.

(2) Data Augmentation is a technique to enhance the size and quality of training datasets so that the deep learning model can be more effectively trained based on the augmented data (Shorten & Khoshgoftaar 2019, 7).



Figure 4. variety of data augmentation (Living Language 2016)

Since the augmented dataset was over 30 GB, therefore, before training the model, I needed to trim down the size of the image from 1536x864 pixel (my laptop screen size) to 640x360 pixel, the final dataset was 5.7 GB. It was much lighter!

#### 4.1.2 How Does Yolo Work with This Dataset?

The goal of Yolo model is to create a bounding box around the detected objects in an image. Yolo divides an image to  $S \times S$  grids (see the image below), with each grid tells if there were any objects in it by the y output. (Redmon & Farhadi 2018, 1-5.)

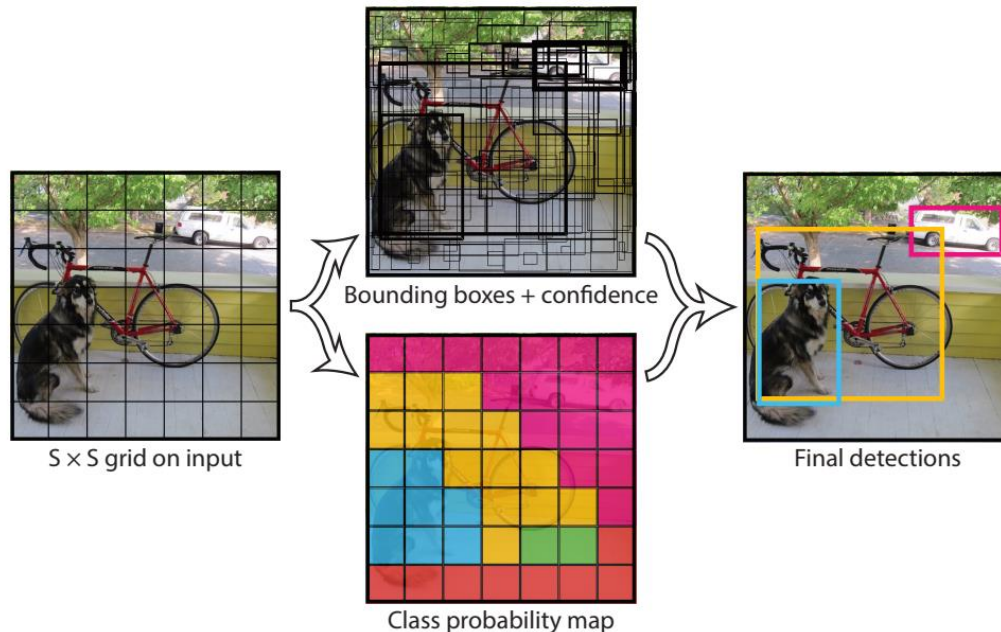


Figure 5. Yolo output (Redmon, Divvala, Girshick & Farhadi 2016, 2)

The y value for each grid for the sign language alphabet looks like this

y =

$p_c$	$b_x$	$b_y$	$b_w$	$b_h$	$c_1$	$c_2$	...	$c_{26}$
-------	-------	-------	-------	-------	-------	-------	-----	----------

- $p_c$ : tells if there is an object in a grid, it is a probability, ranging from the value 0 to 1 (Sharma 2018).
- $b_x, b_y, b_w, b_h$ : tells the position of the bounding box (position and area) in a grid proportionally to the grid size (Sharma 2018).
- $c_1, c_2, \dots, c_{26}$ : tells which class the bounding box belongs to, 26 is the number of the English alphabets, one of these values is assigned with value 1,

corresponding to the detected class, while the others are assigned with value 0 (Sharma 2018).

That is the case when each grid has only one bounding box of an object, what if there are multiple objects in a grid? Object detection models must deal with a concept called anchor boxes, which mean when multiple bounding boxes stack over the others on the same spot of the image. (Sharma 2018.)

The y output of a grid when occurring the anchor boxes looks like this

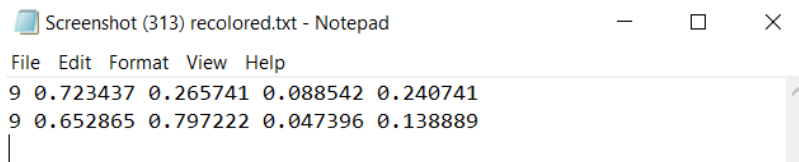
$$y = \underbrace{p_c \quad b_x \quad b_y \quad b_w \quad b_h \quad c_1 \quad c_2 \quad \dots \quad c_N}_{1_{st} \text{ bounding box}} \quad \dots \quad \underbrace{p_c \quad b_x \quad b_y \quad b_w \quad b_h \quad c_1 \quad c_2 \quad \dots \quad c_N}_{k_{th} \text{ bounding box}}$$

That is the case when there are k anchor boxes appearing in a grid, and each anchor box have N+5 values representing it, with N as the number of classes, here I had 26 classes. Usually, in practice, the value of k is chosen to be maximum 3 to train the model effectively.

Combining the information, I had from this section, an image can give an output of size  $S \times S \times (k \cdot (N+5))$ , with S as the grid size to split an image, k is the number of possible anchor boxes each grid has, N is the number of classes of the dataset.

#### 4.1.3 Labeling Tool

Firstly, let's understand what y train label should look like. The format of the y value in training data is [c, x, y, w, h], with c telling the which class the object belongs to, x is the horizontal center of the bounding box, y is the vertical center of the bounding box, w/h is the width/height of the bounding box. The c is an integer value: 0, 1, ..., N, with N is the number of classes in the training dataset; x, y, w, h is the float number ranging from [0, 1], which indicate the position of bounding box proportionally to the image size.

A screenshot of a Notepad window titled "Screenshot (313) recolored.txt - Notepad". The window contains two lines of data: "9 0.723437 0.265741 0.088542 0.240741" and "9 0.652865 0.797222 0.047396 0.138889". The text is displayed in a monospaced font on a white background with a light gray border.

```
Screenshot (313) recolored.txt - Notepad
File Edit Format View Help
9 0.723437 0.265741 0.088542 0.240741
9 0.652865 0.797222 0.047396 0.138889
|
```

Figure 6. Data labelling output

The tool I used to create the y input above is Labelling (Lin 2022), an open-source project available on GitHub, developed by a software engineer named Tzuta Lin. The instruction in the README.md file is very straight forward so I did not cover detailed step here. Basically, the output is the “.txt” file that have the same name as the image name, its content is just as described above. When there are multiple bounding boxes in an image, the content is presented in multiple lines, each of which tells the information of each bounding box.

## 4.2 Training and Evaluation

In this subchapter, I give you a closer look at my training and evaluation process on my Yolo model. You could also find the detailed information of website application I built in the evaluation section.

### 4.2.1 Training

I used Google Colab to train the Yolo model. Google Colab allows users to utilize GPU’s core to train model. This product is used by researchers in several applications, especially in the field of ML and deep learning (Nanda 2020, 42).

- Step 1: clone the darknet repository from Bochkovskiy’s github repository <https://github.com/AlexeyAB/darknet>
- Step 2: to utilize the GPU’s core, set parameter GPU, CUDNN, OPENCV to 1 in Makefile file in darknet repository (Wotherspoon 2020).
- Step 3: copy and modify the cfg/yolov3.cfg file to match the customed dataset. Here I had the dataset of 26 classes to change that file’s content accordingly (Wotherspoon 2020).

- Step 4: create a new data/<projectname>.data file specifying basic information, such as the training and testing image files location, folder to store the output model, total classes, list of class name (Wotherspoon 2020).

I did not mention those steps in details here, a research document. Instead, one could check the YouTube tutorial “YOLOv3 in the CLOUD” by “The AI Guy” channel sited in the references section, this source is officially certified by the author of Yolov4, Profs Bochkovskiy (Wotherspoon 2020).

#### 4.2.2 Evaluation

Instead of collecting another data from other YouTube video, the evaluation is executed by interviewing 10 invited people who did not contribute to the training image. The purpose was to test the model with data of different hands and several backgrounds to see if the model performed well on the unseen data. To organize such research with an optimal cost, I used the Yolov3-tiny model so that it can be deployed with a CPU core of a free webhost, so that I did not need to buy a GPU core server which is not very economical for the survey purpose.

Survey description:

Each interviewee was required to have decent internet connection to access to this website application <https://asl-yolo-detection.herokuapp.com> and a laptop with a camera opened to capture the video of themselves, directly from the browser camera.

Then he/she had the access to a Google survey which contains two parts and has 52 questions in total (26 questions for each part for 26 alphabet characters); the first part is to evaluate the model with the natural light without the support from flashlight; the second part is to evaluate the model with the support from flashlight.

The interviewee has 3-5 trials for each question to get the best result in the web application, then go back to the Google survey form to answer the question.

Question/Answer description:

Each question in the survey has four answers:

- 1) Can't detect (no bounding box generated)
- 2) Detect to a wrong letter
- 3) Can detect the letter but I have to adjust my hand posture
- 4) Detect perfectly (a bounding box generated as soon as my hand appear on the screen)

Answer 1 is chosen when the interviewee tried 5 times but there is no bounding box drawn on all the images. Answer 2 is chosen when the interviewee tried 5 times and a few of which were detected but none of the detected images has the correct alphabet character. Answer 3 is chosen when the interviewee tried 2-5 times and at least one of which had detected the correct alphabet character. The answer 4 is chosen when the interviewee got the correct detection in the first trial or 3/4 or 4/5 trials has the correct detection.

Webapp using description:


The webapp captures an image/frame from the real time video of the interviewee's webcam then send that image/frame to server site to process the image and send it back to client site as small images right below the main camera area. See the image below for more intuitive description of the webapp layout.



Figure 7 (3). The web application layout is used for survey purpose

There are two modes of the webapp: “Main-screen” mode and “Slide” mode.

Using instruction:

- In the “Main-screen” mode (where user can see the camera area and send/receive image from/to the server):
  - Press the Enter ↵ button on keyboard to take a photo (it automatically sends to the server site for processing)
  - Press Delete button on keyboard to delete the latest photo OR press the  button on each photo to delete respectively.
  - Click on the tiny photo to enter the “Slide” mode
- In the “Slide” mode (Where you can see the processed the image clearly):
  - Use the < > button on the screen OR press ⇐ ⇒ on keyboard for sliding to the next/previous photo
  - Press Escape/ESC on keyboard to get back to the “Main-screen” mode

Using rules:

- You have to wait for a photo to be processed completely before taking another one.
- They keyboard manipulation in mode cannot be executed when you are in the other mode. For example: when you in “Slide” mode, you cannot take a new photo by pressing the Enter ↵ button the action of which action belong to “Main-screen” mode.

Result description and discussion:

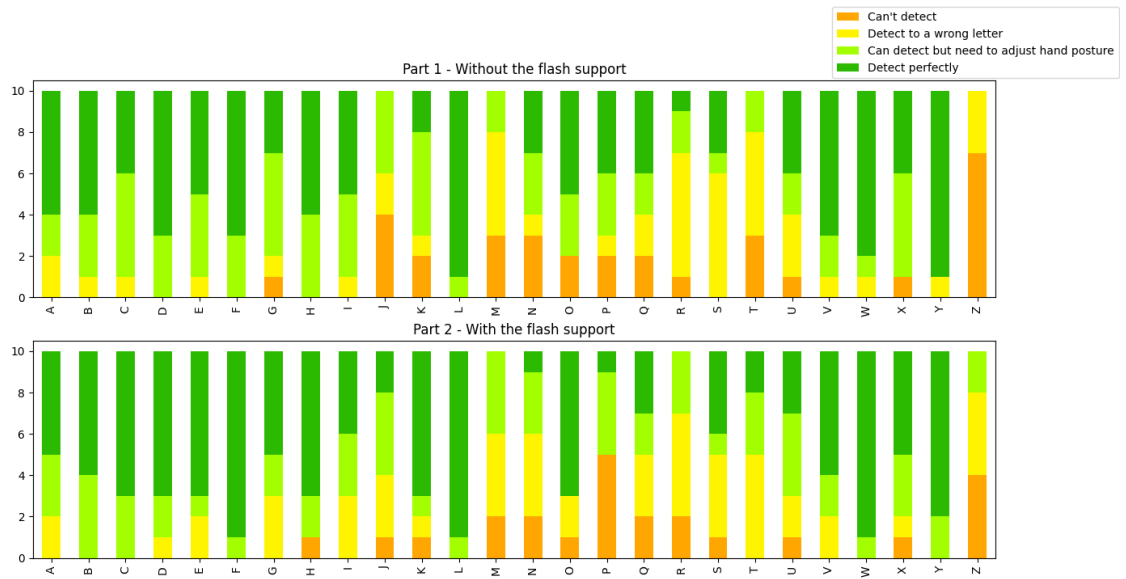


Figure 8. stacked bar graphs to visualize the results of part 1 and part 2 of the survey

The overall result was decently good, the Yolo model was able to detect the untrained data. However, few alphabets were not well detected such as “M”, “N”, “T”, “J”, “Z”, “P”, “Q”, “R”. Among those character, the letter “M”, “N”, and “T” were mis-detected to each other because of its similar posture; “J” and “Z” were motion-based character, both were almost undetectable with the non-flashlight survey; “P” and “Q” were a little tricky to recognize; “R” was observed as the most misinterpreted character. The remaining character was detected decently well despite the variety of background. Finally, you can see a small improvement on the flashlight support part (part 2) in the major of character compared to the non-flashlight part (part 1) because more contrast was created to help the Yolo model distinguish different parts of the hand as well as the hand from the background.

One more observation from the survey is that the plain color background gives a better result than the background which has many objects/details in it. For example, see the examples of the two interviewees below.





Figure 9 (3). The model got confused between letter A and N, because of the detailed background

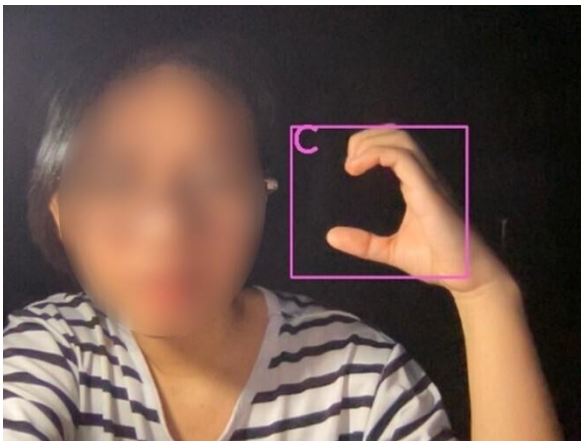


Figure 10 (3). The letter C was detected well thanks to the plain color background

There is a way to improve the performance of the model is to feed more data to train the model with higher variety of backgrounds and even color skins and hand shape. In this project because of the limitation of the resources (the training data mostly comes from internet sources), I had not had ample training data as I wanted, hence, it could be said that Yolo model detect passingly well on the unseen data.

(3) I had been allowed by the interviewees to utilize photo for educational purpose in this document.

## 5. CONCLUSION

In this document, I provided you, probably, with your very first insight about the deaf-and-mute community and sign language, such as the main features constructing sign languages. Besides, you knew more about the current hardship that they are going through with the hearing aids.

In the research and theory part, it was discussed how computer vision, especially Yolo model, deal with sign language and other existing approach to translate the sign language to oral speaking language. I walked you through the basic mathematic concepts constructing Yolo and how they make Yolo special and stand out in comparison to the prior object detection model.

In the construction part. I built an application as a demonstration of how Yolo works and used them for evaluating the performance of the customized model which were trained based on my own uniquely generated dataset. The evaluation result indicated that my Yolo model performed poorly on the dynamic gesture and well on the static posture. My dataset should be more diverse in terms of light condition and backgrounds to train the Yolo more efficiently.

Based on my evaluation result, I concluded that Yolo is not a proper solution for translating the sign language despite of its speed and efficiency in detecting static object because of one of the main features of sign language, the dynamic hand movement. The future technology needs to evolve further to build a two-way communication bridge between the deaf-and-mute and normal speaking community.

## BIBLIOGRAPHY

- ASL That 2013. The ASL Alphabet | ASL - American Sign Language – ABCs. Accessed 29 December 2021. <https://www.youtube.com/watch?v=tkMg8g8vVUo>, YouTube Video.
- Bhowmick, S. & Kumar, A. 2015. Hand Gesture Recognition of English Alphabets using Artificial Neural Network. <https://doi.org/10.1109/ReTIS.2015.7232913>.
- Brown, S. 2021. Machine learning, explained. Accessed 29 January 2022. <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>
- Dellis, N. 2020. MEMORIZE the ASL alphabet (American Sign Language). Accessed 28 December 2021. <https://www.youtube.com/watch?v=E9hu7XGR3-c>, YouTube Video.
- Gopinath, B., Wang, J. J., Schneider, J., Burlutsky, G., Snowdon, J., McMahon, C. M., Leeder, S.R. & Mitchell, P. 2009. Depressive symptoms in older adults with hearing impairments: the Blue Mountains Study. *J Am Geriatr Soc.* Vol 57, Issue 7, 1306-1308. <https://doi.org/10.1111/j.1532-5415.2009.02317.x>.
- Guth, D. 2020. Deaf Awareness Month: 10 Things to Know About Being Deaf. Accessed 26 January 2022. <https://www.hearinglikeme.com/deaf-awareness-month-10-things-to-know-about-being-deaf/>.
- Kochkin, S. 2012. Hearing loss treatment. Better Hearing Institute. [http://www.betterhearing.org/hearing\\_loss\\_treatment/index](http://www.betterhearing.org/hearing_loss_treatment/index).
- Learn How to sign 2020. Learn How to Sign The Alphabet (ABCs) in ASL. Accessed 25 December 2021. [https://www.youtube.com/watch?v=bFv\\_mLwBvHc](https://www.youtube.com/watch?v=bFv_mLwBvHc), YouTube Video.
- Lin, T. T. 2022. labellmg. Accessed 17 January 2022. <https://github.com/tzutalin/labellmg>.
- Living Language 2016. Learn the American Sign Language. Accessed 26 December 2021. <https://www.youtube.com/watch?v=jEB45Z6xIAg>, YouTube Video.
- Loiseau, J. C. B. 2020. Binary cross-entropy and logistic regression. Accessed 1 February 2022. <https://towardsdatascience.com/binary-cross-entropy-and-logistic-regression-bf7098e75559>.
- Lukka, K. 2003. The Constructive Research Method. The Turku School of Economics and Business Administration.
- McCormack, A. & Fortnum, H. 2013. Why do people fitted with hearing aids not wear them? *International Journal of Audiology*, 52:5, 360-368, <https://doi.org/10.3109/14992027.2013.769066>.

McGregor, C. 2018. Using Constructive Research to Structure the Path to Transdisciplinary Innovation and Its Application for Precision Public Health with Big Data Analytics. *Technology Innovation Management Review*.

Nanda, M. 2020. YOU ONLY GESTURE ONCE (YOUGO): AMERICAN SIGN LANGUAGE TRANSLATION USING YOLOV3. Purdue University. Department of Computer and Information Technology. Master thesis.

Pribanić, L. 2006. Sign Language and Deaf Education: A new tradition. *Sign Language & Linguistics*. <https://doi.org/10.1075/sll.9.1.12pri>.

Radecic, D. 2020. Softmax Activation Function Explained. Accessed 1 February 2022. <https://towardsdatascience.com/softmax-activation-function-explained-a7e1bc3ad60>.

Rautiainen, A., Sippola, K. & Mättö, T. 2017. Perspectives on Relevance: The Relevance Test in the Constructive Research Approach. *Management Accounting Research*, 34, 19-29. <https://doi.org/10.1016/j.mar.2016.07.001>.

Redmon, J. & Farhadi, A. 2016. YOLO9000: Better, Faster, Stronger. <https://doi.org/10.48550/arXiv.1612.08242>.

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. 2016. You Only Look Once: Unified, real-time object detection. <https://doi.org/10.48550/arXiv.1506.02640>.

Redmon, J., Farhadi, A. 2018. YOLOv3: An Incremental Improvement. <https://doi.org/10.48550/arXiv.1804.02767>.

Sharma, P. 2018. A Practical Guide to Object Detection using the Popular YOLO Framework – Part III (with Python codes). Accessed 6 February 2022. <https://www.analyticsvidhya.com/blog/2018/12/practical-guide-object-detection-yolo-framework-python/>.

Shorten, C. & Khoshgoftaar, T. M. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*. <https://doi.org/10.1186/s40537-019-0197-0>.

Sign Solutions 2021. What are the different types of sign language? Accessed 28 January 2022 <https://www.signsolutions.uk.com/what-are-the-different-types-of-sign-language>.

Warden, P. 2017. How many images do you need to train a neural network? Accessed 5 February 2022. <https://petewarden.com/2017/12/14/how-many-images-do-you-need-to-train-a-neural-network/>.

World Health Organization 2021. Accessed 28 January 2022. <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>.

Wotherspoon, J. 2020. YOLOv3 in the CLOUD: Install and Train Custom Object Detector (FREE GPU). The AI Guy. Accessed 7 January 2022. <https://www.youtube.com/watch?v=10joRJt39Ns>, YouTube Video.

Yuan, Y. 2020. YOLO Creator Joseph Redmon Stopped CV Research Due to Ethical Concerns. Accessed 30 January 2022.  
<https://syncedreview.com/2020/02/24/yolo-creator-says-he-stopped-cv-research-due-to-ethical-concerns>.