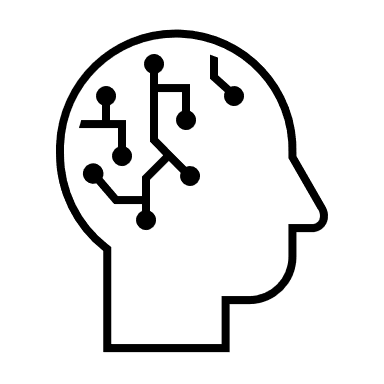
**FINAL REPORT**

UTILISING DEEP LEARNING IN SIGN LANGUAGE RECOGNITION & INTERPRETATION

**AUTHORS:**

CHAN JOSHUA JUAN YIN

BRANDON LIN ZHAN HONG

DUONG NGOC YEN

PHAM VAN LONG PHUOC

YU XINHUI

WANG YUJING

**DATE OF SUBMISSION:** 9th MAY 2021

**Table of Contents**

[1. Background & Motivation 3](#_Toc71308174)

[2. Product Showcase 4](#_Toc71308175)

[2.1. Features 4](#_Toc71308176)

[2.1.1. Near-Real-Time ASL Interpretation 4](#_Toc71308177)

[2.1.2. Graphical User Interface (GUI) 4](#_Toc71308178)

[2.2. Value Proposition 5](#_Toc71308179)

[2.2.1. Advantages 5](#_Toc71308180)

[2.2.2. Disadvantages 5](#_Toc71308181)

[2.3. Future Improvements 6](#_Toc71308182)

[2.3.1. Applicability with Natural Language Processing (NLP) 6](#_Toc71308183)

[2.3.2. [Insert] 6](#_Toc71308184)

[3. Approaches & Outcomes 7](#_Toc71308185)

[4. Project Takeaways & Reflections 10](#_Toc71308186)

[4.1. Chan Joshua Juan Yin 10](#_Toc71308187)

[4.2. Brandon Lin Zhan Hong 10](#_Toc71308188)

[4.3. Duong Ngoc Yen 11](#_Toc71308189)

[4.4. Pham Van Long Phuoc 11](#_Toc71308190)

[4.5. Yu Xinhui 12](#_Toc71308191)

[4.6. Wang Yujing 12](#_Toc71308192)

[ANNEX A: Individual Contributions 13](#_Toc71308193)

[ANNEX B: References 14](#_Toc71308194)

[ANNEX C: [Insert anything if necessary] 15](#_Toc71308195)

1. Background & Motivation

The ability to hear is often taken for granted. However, auditory perception is not universal trait. The World Health Organisation (WHO) estimates that more than 5% of the world’s population – approximately 466 million people – have some form of disabling hearing loss, and this number is estimated to grow to in excess of 900 million people by 2050 [1].

In lieu of verbal communication, members of Deaf cultures use sign languages as their primary means of communication within their communities. Meaning is conveyed through a combination of hand shapes, hand orientation, and the movement of one’s hands, arms and/or body, as well as one’s facial expressions. Due to the distinct difference in communication mediums between hearing people and members of Deaf cultures, effective interaction is not straightforward. As human society consists primarily of hearing people, it is generally difficult for a hearing person and a deaf person to converse without the aid of a sign-language interpreter or technological accessibility, such as through texting. Members of the Deaf community may thus feel misunderstood or be discriminated against by those who do not understand sign language [2].

There have been attempts made in the past few decades to overcome this unique communication barrier. The introduction of cochlear implants has helped many users gain auditory perception to a certain degree, but their use is not widespread and has faced resistance from members of Deaf communities, requiring time to become more widely accepted and utilised. On the other hand, with regards to the use of text messages as an alternative means of communication, this method merely forces communication through a secondary medium, at the expense of the Deaf person’s convenience.

2. Product Showcase

In order to overcome the communication barrier, this Team believes a better approach to ameliorating this problem is to enable hearing people to better comprehend sign language. We seek to do so via a mobile application which, via the use of deep learning models capable of interpreting American Sign Language (henceforth referred to as ASL) in near real-time and forming coherent sentences through Natural Language Processing, will be able to interpret a video input of communication in ASL into American English.

2.1. Features

2.1.1. Near-Real-Time ASL Interpretation

2.1.2. Graphical User Interface (GUI)

2.2. Value Proposition

2.2.1. Advantages

2.2.2. Disadvantages

2.3. Future Improvements

2.3.1. Applicability with Natural Language Processing (NLP)

2.3.2. [Insert]

3. Approaches & Outcomes

Our product development can be separated into 3 Phases:

1. Data Preparation
2. Model Training
3. Application Development

In the process of development, we experimented with different methods to obtain a suitable deep learning model to interpret ASL hand gestures into separate English words. The following, in chronological order, are the various Phases of Development we went through, with Roman Numerals (I, II, III, IV, etc.) used as indication of our nth attempt.

**Data Preparation I**

**Approach:**

We extracted the hand features as NumPy arrays using MediaPipe, a framework developed by Google Research in 2019 [3] used to provide real-time body, facial and hand tracking and extract important landmark features.

**Outcome:**

We were able to obtain word-level hand features in Four-Dimensional (4D) NumPy array for use with two hands. The four dimensions are: *number of frames*, *two hands, 21 hand landmarks per hand per fram*e, and *3 dimensions for each landmark (x, y, depth)*. Hence, the shape of the array is [n, 2, 21, 3] where n is the number of frames of a video.

**Model Training I**

**Approach:**

We tested three different models, *Multilayer Perceptron (MLP) Classifier*, *Long Short-Term Memory (LSTM)* and *Gated Recurrent Unit (GRU)*, using a very small subset of **Data Preparation I**.

* **MLP Classifier:** A type of feedforward artificial neural network (ANN). MLPs are useful in research for their ability to solve problems stochastically, allowing for approximate solutions for extremely complex problems such as fitness approximation.
* **LSTM:** A type of recurrent neural network (RNN) model, developed to deal with the vanishing gradient problem. They are well-suited to classifying, processing, and making predictions based on time series data.
* **GRU:** Another type of RNN model which targets the problem of vanishing gradient during backpropagation. Compared to LSTM, it is newer and uses fewer parameters in order to increase the speed of training.

**Outcome:**

Based on our small subset test, predictions using the MLP model were very unstable. In contrast, predictions using the GRU model were stable and achieved the highest level of accuracy overall. Combining our online research and this trial result, we decided that GRU would produce the most suitable model.

Our group discovered that there were NaN (Not a Number) values in the dataset. These arrays containing NaN values, corresponding to frames in which the hands did not appear, result in computations with NaN outputs, confusing the model and thus reducing model accuracy.

Another problem faced was that the video files we utilised had varying numbers of frames, resulting in non-homogeneity in dimensions, resulting in errors in NumPy manipulations.

**Data Preparation II**

**Approach:**

In order to resolve the issue caused by the NaN values, we replaced all NaN values with zeroes. Our initial thoughts were that because of the characteristics of GRU, a frame consisting zeroes, as opposed to NaN values, would not have a significant impact on the model prediction.

**Model Training II**

**Approach:**

Utilising Data from **Data Preparation II**, we re-trained the GRU model after padding the arrays with zeroes so that each array representing a video file contains 233 frames, homogenising dimensions.

**Outcome:**

The model achieved an accuracy around 3%. We believe this is due to a large number of insignificant frames (padded by zeroes), as well as the inherent bias in the dataset from certain words having few corresponding videos while others having multiple. This resulted in more unreliable predictions for less-represented words.

**Data Preparation III**

**Approach:**

Based on the model from **Model Training II**, each array consisted of 233 frames, with zeroes present. We decided to exclude frames with zeroes to eliminate their impact.

We noted that processing all 233 frames in each video file would result in significant repetition, slowing down the model training process. Hence, we sampled randomly 16 non-zero frames for each video, with repetitions allowed for assured randomness.

To mitigate data bias, we dropped words with less than 16 corresponding videos such that the model could predict more accurately.

**Outcome:**

Data from **Data Preparation III** contained hand features for 111 words. Each of these words have more than or equal to 16 corresponding videos. Each video was represented by a 4D array in the shape [16, 2, 21, 3].

Model Training 3.0

Approach: Using data 3.0, we trained the GRU model again and predicted the words with top 3 possibility and top 10 possibility respectively. We then calculated the accuracy of prediction.

Outcome: The train set accuracy was 97.46%. In the test prediction, the accuracy was16.23%. There was a 34.34% chance the correct word was among the top 3 probable words and a 61.89% chance that the top 10 probable words included the correct word.

Model Training 4.0

Approach: As we trained the GRU model using Data 3.0, the same 16 frames for each video were used, which meant that some frames of a video was never utilized to train the model. Thus, we decided that random sampling of the 16 frames should occur at the beginning of each epoch. This is to ensure the model can learn from a different set of 16 frames each time.

Outcome: Using this approach, we managed to achieve a train accuracy of 91.33%, a test accuracy of 29.06%, 48.30% for top 3 probable words and 73.96% for top10 probable words.

Comparison of Different Types of RNN Models

The above graph is the change of accuracy with respect to epochs, using three different types of model: RNN, GRU, LSTM. GRU was proven the best model amongst the three.

Comparison of GRU Model Accuracy using different no. of frames

The above graph is the change of accuracy with respect to epochs, when the model is trained using different numbers of frames for each video. From the graph, we could see that 8 frames per video allowed for highest accuracy. We thus decided to use 8 frames per video for training of the GRU model.

Word Labels

Since there are a total of 1970 different words in our dataset, we used 1970 numbers as indexes for words and created an index:word label dictionary for display of our final translation.

4. Project Takeaways & Reflections

4.1. Chan Joshua Juan Yin

4.2. Brandon Lin Zhan Hong

4.3. Duong Ngoc Yen

4.4. Pham Van Long Phuoc

4.5. Yu Xinhui

4.6. Wang Yujing

ANNEX A: Individual Contributions

|  |  |
| --- | --- |
| CHAN JOSHUA JUAN YIN | * A * B * C * D * E |
| BRANDON LIN ZHAN HONG | * A * B * C * D * E |
| DUONG NGOC YEN | * A * B * C * D * E |
| PHAM VAN LONG PHUOC | * A * B * C * D * E |
| YU XINHUI | * A * B * C * D * E |
| WANG YUJING | * A * B * C * D * E |

ANNEX B: References

ANNEX C: [Insert anything if necessary]