WALDO

Watson Assisted Living Device – WALDO

Joshua Rizal Chan, Lua Ying Hao, Ng Yi Song, Patrick John Chia, Yeow En Kai Joel

Department of Electrical and Electronic Engineering, Imperial College London

Imperial College London

IBM

What is WALDO ?

- A cute and endearing device that performs low latency Makaton Sign Language Recognition and Vocalisation using Machine Learning
- Contains on-device buttons that the user can press to vocalise pre-set phrases in a quick and convenient manner
- Has customisable on-board ambient sensors to help the carer or medical professionals to understand the environmental conditions of the user

Why WALDO ?

- Care homes for people with learning difficulties face high staff turnover and staffing costs
- Training staff to use Makaton is expensive and time-consuming
- WALDO aims to provide a streamlined, cost-effective way for care homes to provide high-quality care for Makaton users by easing communication

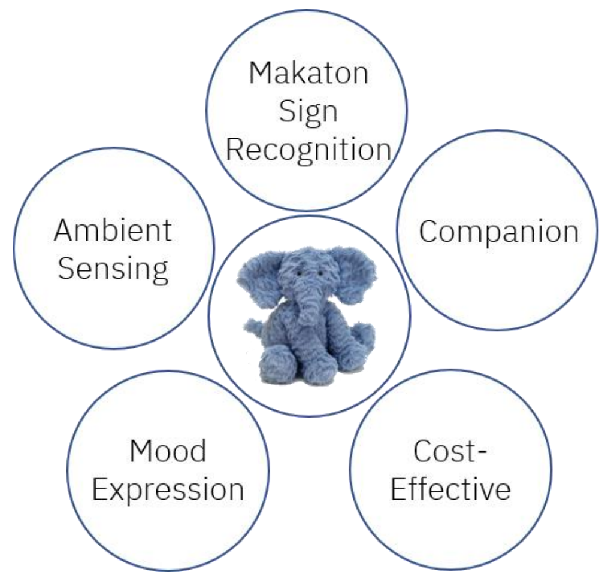

Figure 1: WALDO's core capabilities

Methodology

Choice of Makaton Signs

- 5 Makaton signs chosen based on the following considerations:
  - Recommendations from Precious Homes, a care home for people with learning disabilities and other needs
  - Static and moving signs to demonstrate device capabilities

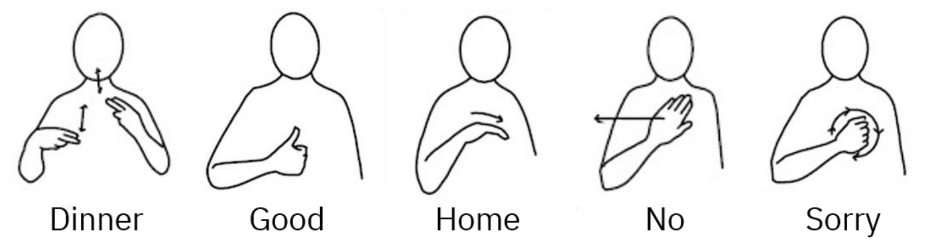

Figure 2: 5 Makaton signs chosen for implementation

- Volunteers[1] found around the Imperial College South Kensington campus – videos of volunteers performing Makaton signs were recorded for dataset

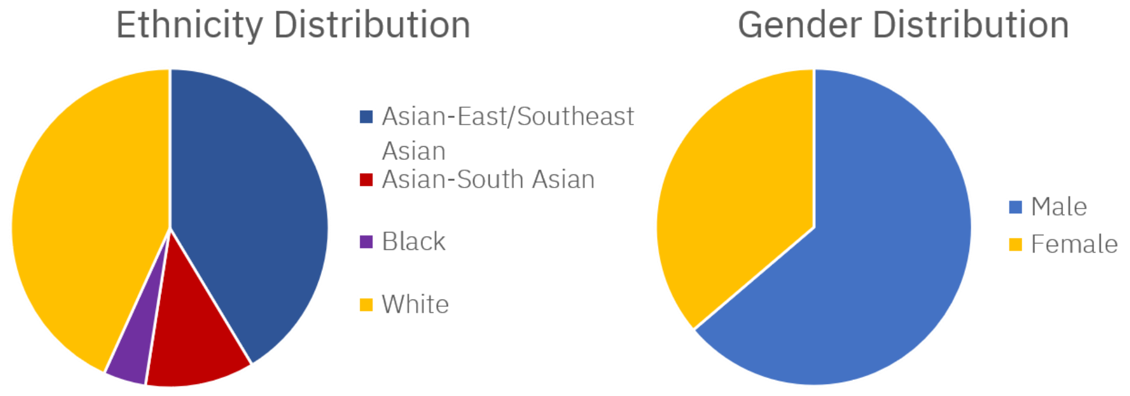

Figure 3: Demographic distribution of training dataset

- The diversity of the training dataset was maximised to ensure robustness when used by different users

Machine Learning employed to interpret these Makaton signs

[1] Volunteers in this project provided their consent to the video recordings on the basis that the recordings were only to be used for this project and would be deleted at the end of the project.

Makaton Sign Language Recognition with Machine Learning

Classification Problem : 6 Classes –– 5 Makaton Gestures + 1 No Action Class

Input : 30 contiguous frames, each of dimension 640x480x3 pixels, $x \in \mathbb{R}^{30\times640\times480\times3}$

Hypothesis : $h \in H$ where H is the set of possible hypothesis

Output : $\hat{y} = h(x) = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_6]^T \in \mathbb{R}^{6x1}$, predicted probability of each class $\hat{y}_n$ given input x

Loss Function : Categorical Cross Entropy, to heavily penalise confident but wrong predictions.

$$l[h(x)] = -\sum_{c=1}^{c=6} y_c \log \hat{y}_c$$

where y_c is binary (0 or 1) and indicates if class c is the correct classification given observation x.

Aim : Obtain best model g such that

$$g = \underset{h\in H}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} l[h(x_i)] \right\}$$

where N is the size of the training set with x_i the inputs from the training set.

Deep Neural Network Architecture

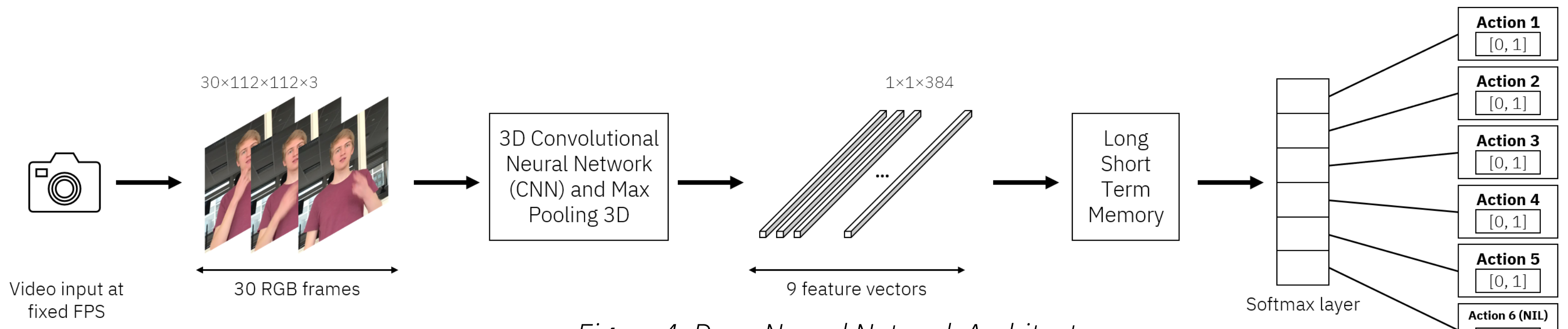

Figure 4: Deep Neural Network Architecture

- Important to capture spatio-temporal information for accurate gesture recognition
- Proposed architecture achieves this by using both 3D Convolution (Conv3D) and Long Short Term Memory (LSTM)
- Conv3D performs convolution not only along the height and width domain but also the temporal domain of the input frames – both spatial and temporal features learned and passed into the LSTM
- LSTM further studies the time-evolution of the spatio-temporal features captured before making a decision on the gesture recorded

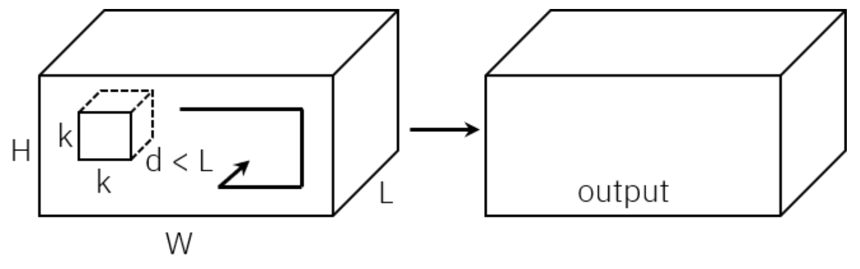

Figure 5: 3D Convolution [2]

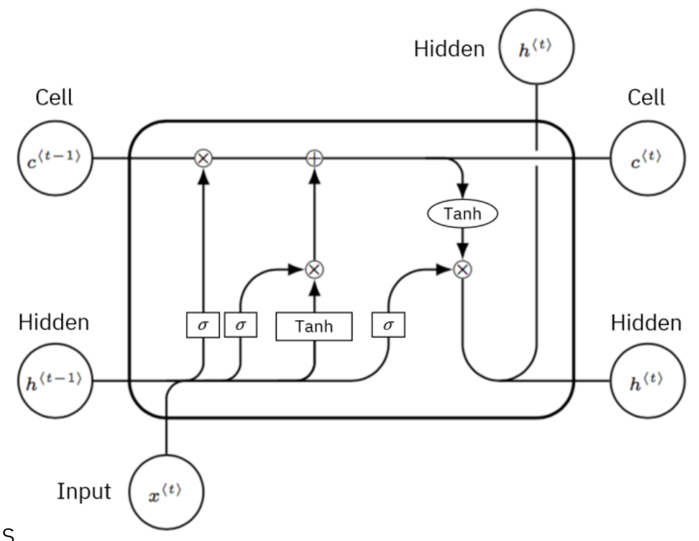

Figure 6: LSTM Cell

[2] Tran, D., Bourdev, L., Fergus, R., Torresani, L., et al. (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. *2015 IEEE International Conference on Computer Vision (ICCV).*

Model Performance

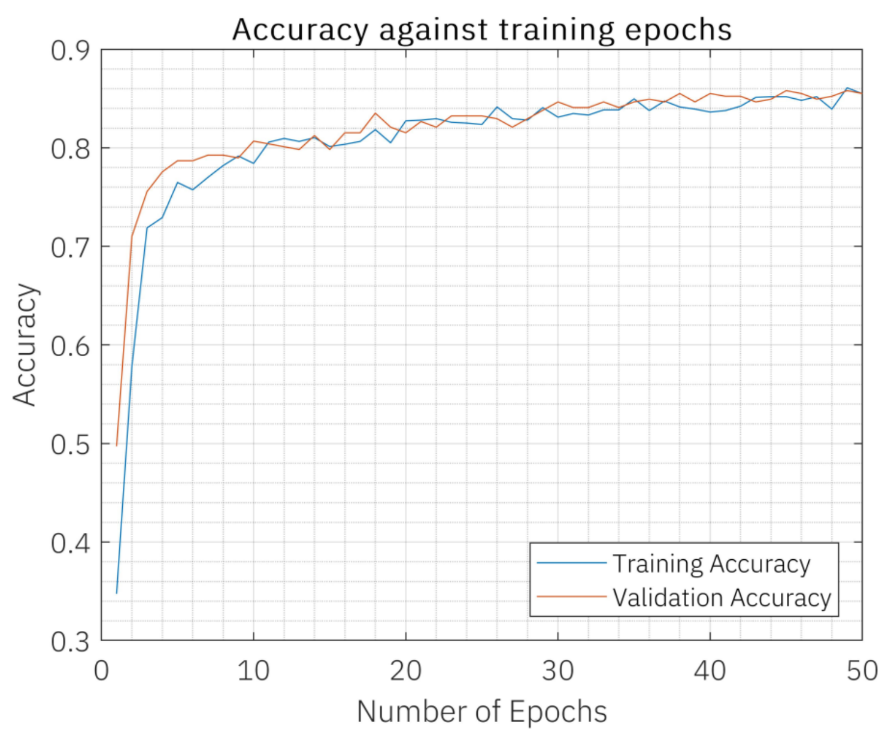

Figure 7: Evolution of model accuracy

Model Evaluation

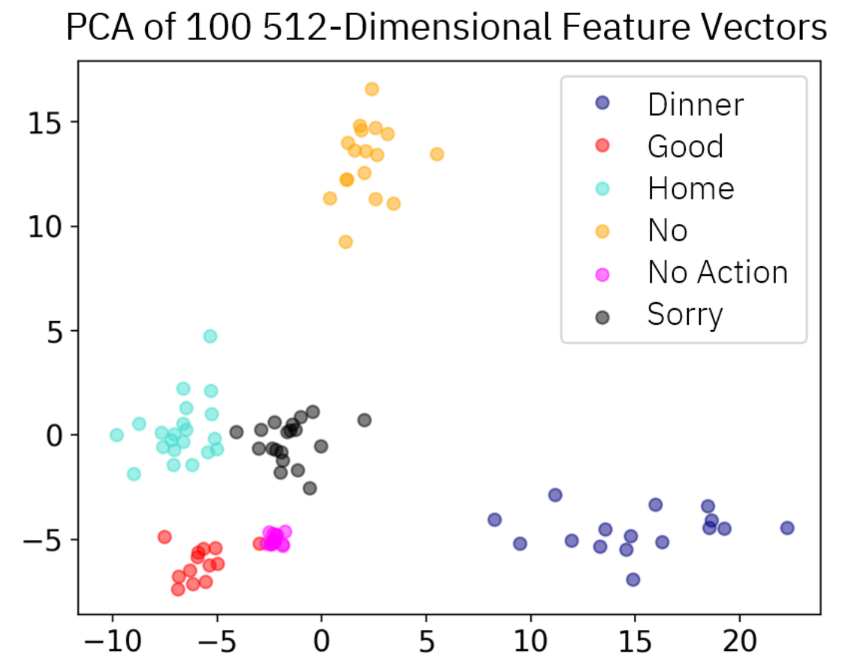

Figure 8: Principle Component Analysis (PCA) plot of compressed feature vectors

- Visualisation of 512-dimension vector space with 2 dimensions using PCA
- Clear clustering of each class can be observed
- Distance between features from actions Home, Good & Sorry closer than desired, but explainable given innate similarity

WALDO Hardware

- Machine learning implemented on Nvidia Jetson Nano – contains GPU optimised for running convolution operations required in machine learning
- To limit computational overhead on the Jetson Nano, a Raspberry Pi performs all other device functions – buttons, speakers and sensors (excluding camera)
- IBM Watson Text-to-Speech implemented on Pi, with output via connected speakers
- Jetson and Pi connected using GPIO pins
  - Pi → Jetson: Ultrasonic sensor information, to start machine learning detection of Makaton signs
  - Jetson → Pi: Action detected by machine learning model
- Power source: Portable power bank used as a power source to ensure portability while fulfilling laptop-grade requirements of Jetson

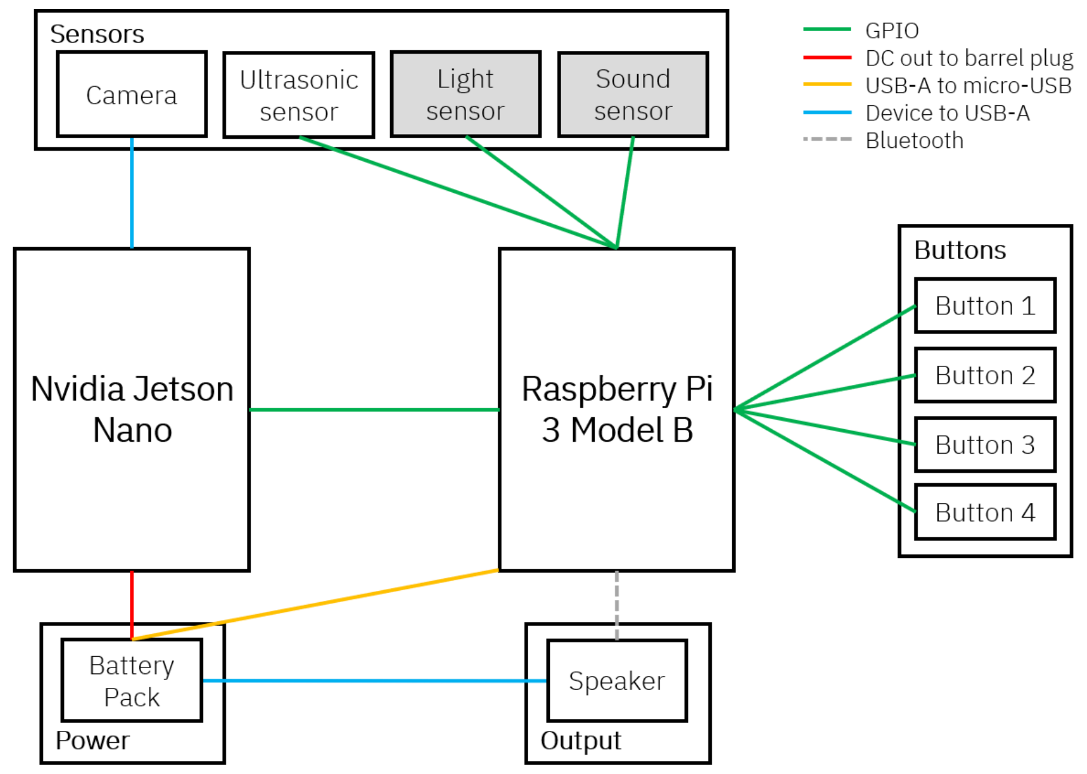

Figure 9: Diagram of device hardware and connections (gray sensors not implemented)

Additional Considerations

- **Network compressed** sufficiently for edge computing on Jetson
  - Reduced latency (~1.0s)
  - 10x speed up in throughput (~9Hz)
  - To accommodate GPU RAM limitations – only 2.5GB available
- **Data Augmentation** enables sign recognition from anywhere in the frame

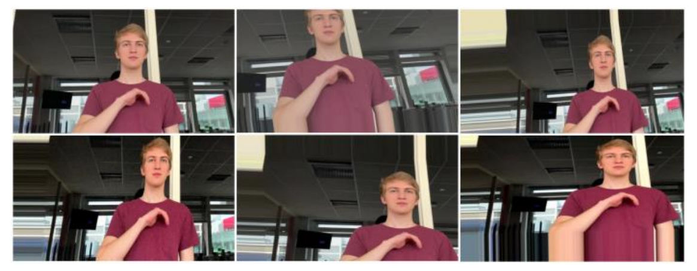

Figure 10: Example of data augmentation (original at top left, all other images randomly modified)

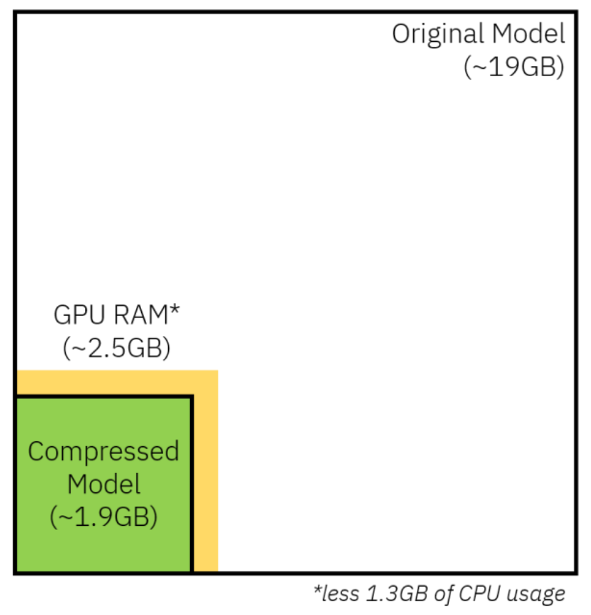

Figure 11: Visualisation of effect of model compression on RAM usage

Conclusion

- The main capability of WALDO, the interpretation of Makaton signs, has been successfully implemented with low latency
- Machine learning for sign language interpretation is feasible and shows great potential for further expansion
- Given more time, WALDO's capabilities can be extended, and its present hardware streamlined, to provide a more holistic system of care for its user

Future Work

- Makaton
  - Increase the number of signs the device can recognise
  - Implement broader-scale, institutionally supported method of data collection to increase size and diversity of dataset
- Device sensors
  - Integrate ambient sensors that provide carers with a fuller understanding of the user's environment, to facilitate better detection of possible risk factors
  - Create a plug-and-play system for sensors to make WALDO an easily upgradeable and expandable system as user needs adapt
- Improve manufacturability of product by designing suitable physical components
- Use purpose-built electronic components to improve performance and power consumption

Acknowledgements

We would like to thank:

- Dr Steve Wright, for his invaluable guidance and encouragement throughout the project
- Louise Cooper, John McNamara & Emily Larkin from IBM for their clear direction and unwavering support in the implementation of this project
- Precious Homes, for supporting this project by providing valuable insight into the use of Makaton, as well as training videos
- Vic, May & Amine at the teaching labs, for their untiring technical support
- All who volunteered their time to help us in the collection of training data