THE CANADA SCIENCE FAIR JOURNAL

ARTICLE

Machine Learning

Highly Evolved Lip Learning mOdel (HELLO): A Novel Machine Learning Model for Accurate Lipreading

Written By: Nimish Jain
Edited by: Tristan Montoya
Layout By: Darren Wang

Awards:

Gold at the Greater Vancouver Regional Science Fair

Bronze at the Canada-Wide Science Fair

Communication is one of the most essential prospects of life and yet over 18 million Americans and 500,000 Canadians with speech impairments lack access to an effective communication method (Black et al., 2015) and (CDAC, n.d). Unable to describe their feelings and thoughts, they can have a considerably hard time navigating daily life, leading to distress. Additionally, these people are often left out due their inability to socialize, adversely impacting their psychological condition and physical being (Page et al., 2022). Current solutions developed to address this issue are slow, rendering them commercially unviable. For instance, ASL (American Sign Language) requires prior working knowledge of the method. Many individuals simply do not have the time or willingness for this undertaking. Due to its ineffectiveness, people who communicate using sign language are often deprived of critical health information (Hoglind, 2018). Other assistive devices such as speech generating devices face the same problems. They involve custom equipment, most of which is manufactured on request. It can also take months to fully integrate into a working solution.

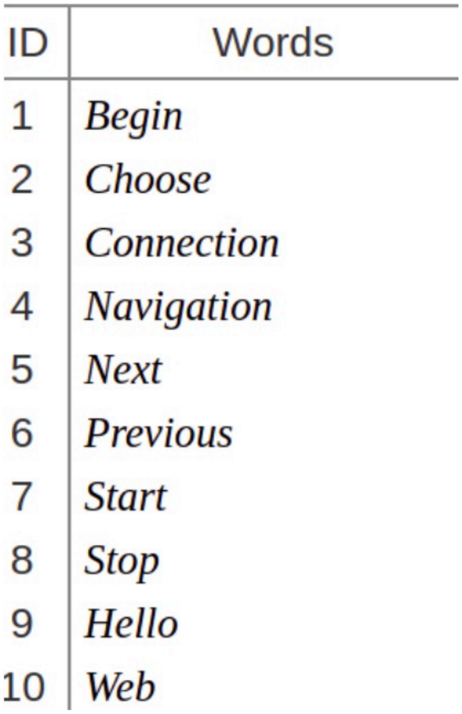
This work is mainly motivated to give these people a voice by providing a free, accessible and easy-to-use solution. By employing a variety of novel machine learning techniques, HELLO analyzes the lip movements in a given soundless video and converts it into text and a playable audio file.

Materials and Methods

HELLO is made up of four parts: the dataset, preprocessing pipeline, machine learning model and web app. The web app is compatible with both mobile and desktop devices and is specifically designed for ease-of-use. With the majority of the global population having access to a smartphone, HELLO is a commercially available solution.

Dataset

Every neural network requires a dataset to train on. Due to the lack of a publicly available dataset for this project, a custom dataset was compiled using videos of myself. The dataset consists of 730 videos with approximately 75 videos for each vocabulary word (*class*). Refer to Figure 1. for a list of vocabulary words used in the dataset. Each video is 25 frames long. A Python script was used to open a video-recording window, where a word was mouthed in the middle of the video, similar to how a person with speech impediments would. It was then saved with its corresponding label. The videos were taken in different lighting conditions and each word was recorded with multiple enunciations to mimic emphasis on certain phonemes in the word. The dataset was split in a 85:15 ratio of training to validation data. The training data was further split in a 90:10 ratio for training and testing. The validation set is used to evaluate model accuracy after training is completed.



ID	Words
1	*Begin*
2	*Choose*
3	*Connection*
4	*Navigation*
5	*Next*
6	*Previous*
7	*Start*
8	*Stop*
9	*Hello*
10	*Web*

Figure 1. Chart containing pairs of vocabulary words and corresponding classes.

Preprocessing Pipeline

This pipeline ensures that only the most relevant parts of the video (regions of interest) are being fed into the model. Before feeding the videos into the pipeline, we apply various augmentations such as scaling, cropping, stretching and distortions. These aid the model in adapting to a wider range of test scenarios. The RGB images are also normalized to the [0-1] scale. The pipeline is a multipart algorithm. First, we split the video into 25 individual frames. For each frame, we first apply an image segmentation model to mark the coordinates of lips. Next, we crop the lip region based on these coordinates and resize the resulting image to 64x64 pixels. After converting to grayscale, this is fed into a convolutional neural network (CNN). This network extracts the *spatial* features of an image (Albawi et al., 2017). These are shapes and lines that uniquely define that image. I use the DenseNet-121 (Huang et al., 2016) for this task. The main advantage of this CNN is the use of residual blocks, which increase speed and result in better features. The resulting matrix of numbers obtained from these processes is called a *feature map*. For a single video, we get 25 individual feature maps which are then concatenated to represent input information for a single video.

Machine Learning Model

I created a modified version of the Transformer architecture (Vaswani et al., 2017) for my model. Transformer models are well-equipped to deal with sequential data and provide immense accuracy and speed gains over traditional architectures. They are divided into three major components: embedding, encoder and decoder layers. The goal of the embedding layer is frames that are spatially similar to each other will have similar representations. A step known as positional encoding is applied to the embeddings. These encodings provide a context for the model to understand where a frame belongs in relation to others. The embedding layer encompasses the output from the DenseNet in the preprocessing pipeline. The eEncoder learns frame-by-frame how these sequences correspond to a given word. It uses a principle called *Attention* to focus on the important frames that have key details. The decoder generates a label based on these learnt representations. The model expresses its output through a *softmax* function (Bridle, 1989). This function tells the model the most probable label a video belongs to.

With the key components ready, the model is trained for a total of 5 *epochs*. An epoch is defined as a complete iteration over all of the training videos. During each epoch, the Transformer model trains on the training dataset and then applies its understanding on the test dataset. After an epoch is completed, an algorithm called back-propagation is applied. It traverses through each of the layers and calculates the error between the true predictions and model predictions. This is known as the *loss*. It then feeds this value to the optimizer, which adjusts the model's parameters (*weights*) to better match the dataset and decrease the loss. At the end of all epochs, a separate algorithm saves the best weights obtained by the model.

Web App

The web app provides a means for the user to utilize the model on any individual video. It is worth noting that the model can only predict the 10 vocabulary words it was trained on. The end-to-end process of the web app is shown in Figure 2. Overall, this process only takes 2-4 seconds, making it very efficient.

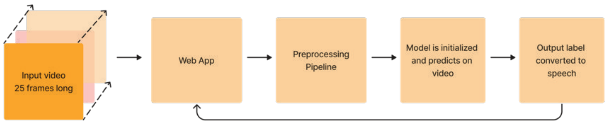

Input video
25 frames long
Web App
Preprocessing Pipeline
Model is initialized and predicts on video
Output label converted to speech

Figure 2. Flowchart illustrating user flow on web app during prediction.

Results

The final model achieves an accuracy of 95.69% with a loss of 0.21 on the training subset. On the test subset, it achieves an accuracy of 92.70% with a loss of 0.25. On the validation set, we get a 91.77% accuracy with a loss of 0.29. Refer to Figures 3 and 4 for loss and accuracy graphs. After initial training, the model was optimized through a tuning algorithm, achieving these final accuracy-loss figures.

S

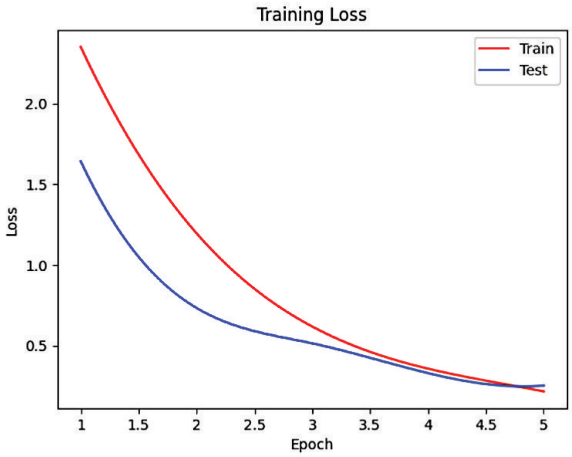


Figure 3. Loss of the model on Train and Test sets with respect to epochs.

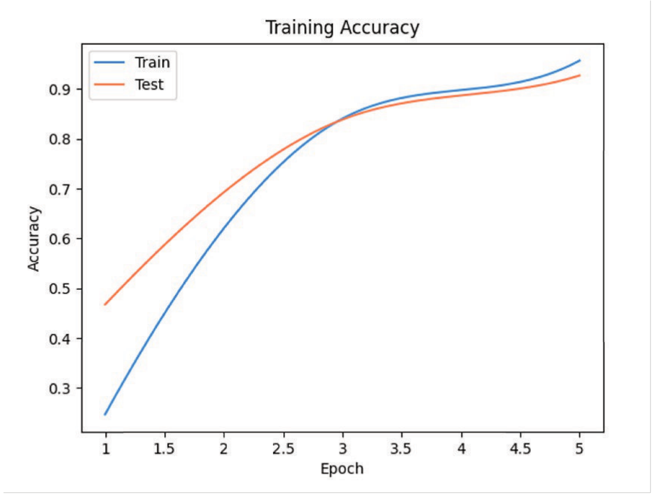


Figure 4. Accuracy of the model on Train and Test sets with respect to epochs.

Discussion

There have been several attempts to establish effective communication methods for the speech impaired but all exhibit drastic limitations. HELLO provides an easy, fast and accessible way for hassle-free communication. With an accuracy of 95.69%, these results show a reliable and fast way to map a video to its corresponding vocabulary word. Apart from that, the high accuracy is achieved in only 5 epochs, showing the effectiveness of the Transformer model. This technology serves as a cure for chronic diseases such as aphonia (inability to produce sound) or dysphonia (incomprehensible speech). Patients with medical illnesses related to speech and voice can use HELLO as an aid in finding employment and accelerating social functioning. Moreover, it can cut down the enormous distress felt after one discovers their inability to speak. Apart from that, it can be used as an effective learning aid for children with special needs inside schools. It can also easily be integrated into modern video conferencing apps in the form of an extension to provide wider range of access. In the future, HELLO will be trained on a larger dataset containing videos of different people with more vocabulary words. Additionally, a website will be created that allows users globally to donate their data.

Conclusion

This project successfully achieved its goal of developing accurate word-level lip reading models. HELLO achieves an accuracy of 95.69% during internal testing, providing a fast and accessible way to recognize speech in silent videos. By implementing novel machine learning techniques and creating a web app, this approach eliminates issues regarding speed and scalability. These issues are faced by current work such as sign language, limiting their efficiency. Apart from proving a way for people with speech impediments to communicate effectively, HELLO can serve as a learning aid.

References

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. International Conference on Engineering and Technology (ICET). https://doi.org/10.1109/ICEngTechnol.2017.8308186

Bhadauria, R., Nair, S., & Pal, D. K. (2007). A Survey of Deaf Mutes. Medical Journal, Armed Forces India. https://doi.org/10.1016/S0377-1237(07)80102-X

Black, L. I., H, M. P., Vahratian, A., & Hoffman, H. J. (2015). Communication Disorders and Use of Intervention Services Among Children Aged 3–17 Years: United States, 2012. National Center for Health Statistics Data Brief. http://www.cdc.gov/nchs/products/databriefs/db205.htm

Bridle, J. S. (1989). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. NIPS. https://dl.acm.org/doi/10.5555/2969830.2969856

CDAC. People who have communication disabilities. Communication Disabilities Access Canada. Retrieved June 12, 2023, from https://www.cdacanada.com/resources/communication-disabilities/statistics/

Chollet, F. (2017). Deep Learning with Python. Manning Publications.

Hoglind, T. A. (2018, October 11). Healthcare Language Barriers Affect Deaf People, Too. Boston University. Retrieved June 12, 2023, from https://www.bu.edu/sph/news/articles/2018/healthcare-language-barriers-affect-deaf-people-too/

Huang, G., Liu, Z., Maaten, L. V. D., & Weinberger, K. Q. (2018). Densely Connected Convolutional Networks. IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.1608.06993

Page, A. D., & Yorkston, K. M. (2022). Communicative Participation in Dysarthria: Perspectives for Management. Brain Sciences 12(4). https://doi.org/10.3390/brainsci12040420

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. NIPS. https://doi.org/10.48550/arXiv.1706.03762

ABOUT THE AUTHOR NIMISH JAIN

Nimish is interested in programming and inventing new things. By following the first principles method, he enjoys learning through experimentation. He is also passionate about writing, sketching and finding new ideas to help people. His current projects include a speech synthesis application to aid people with speech impairments. Apart from programming, Nimish finds immense joy in the physics of airplanes, especially fighter jets. He is filled with curiosity, always eager to learn and ask questions.