Isolated Sign Language Recognition

Chunhao Li
University of Rochester
cli79@u.rochester.edu

Junfei Liu
University of Rochester
jliu137@u.rochester.edu

Abstract

Every day, 33 babies are born with permanent hearing loss in the U.S. Around 90% of them are born to hearing parents, many of which may not know American Sign Language. Without sign language, deaf babies are at risk of Language Deprivation Syndrome. To help connect deaf children and their parents, the Isolated Sign Language competition is launched to promote a feasible and accessible model to provide feedback for sign language practices. We will explore solutions using TensorFlow Lite to build deep learning models capable of running on mobile devices to recognize American sign language in videos.

1 Introduction

1.1 Background

According to the American Society for Deaf Children, 33 babies born in the United States every day have permanent hearing loss, and 90% of these children's parents do not have any background knowledge about American Sign Language. Without access to Sign language, these children with hearing loss might end in Language Deprivation Syndrome(LDS), which means they may not develop the necessary skills to assimilate into the academic learning environment successfully[3].

To help children avoid LDS, Google created a smartphone game app called Popsign and launched the Isolated Sign Language Recognition competition on Kaggle to help them improve Popsign. Participants will build machine-learning models to recognize users' body movements and match them with specific signs in the American sign language with time and space limitations with respect to mobile devices' performance. This model will let Popsign detect users' body movements and sign the corresponding sign language vocabulary, which will help both deaf children and their parents to practice sign language[3]. We propose three approaches to training this task-specific model under strict constraints of time and space in this research project.

1.2 Information about American Sign Language(ASL)

Sign languages are natural languages used by deaf and hard-of-hearing people to communicate with each other and with hearing people who know the language. Sign languages have their own unique grammatical rules, vocabulary, and syntax and are not just a visual representations of spoken languages. Like verbal languages, sign languages are complete and complex languages in their own right, with their own grammar and structure. There are various sign languages that exist in the world. Among several sign languages used in the United States, American Sign Language (ASL) is the primary sign language used by the deaf and hard-of-hearing community in the United States and Canada[3].

ASL is a visual language that uses a combination of hand gestures, facial expressions, and body language to convey meaning. Signs in American sign language and some other sign languages are

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

more like Chinese characters than alphabets; they represent concepts instead of letters, which makes each concept in these sign languages unique. Its signs are organized into five basic parameters: handshape, movement, location, orientation, and non-manual markers (facial expressions and body language). Figure 1 shows the body movements representing the word abandon in American sign language.

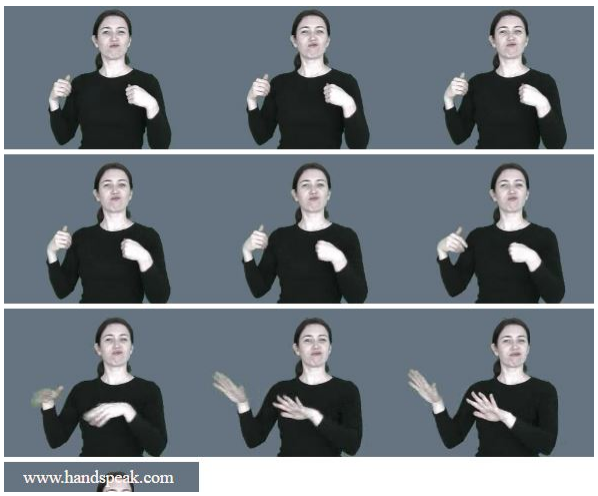

Figure 1: the word Abandon in American Sign Language[10]

ASL is a complex and nuanced language, and learning it requires time, dedication, and practice. Therefore, learning American Sign Language is difficult for parents who are eager to communicate with their deaf children. They want to learn sign language, but it's hard when they are working long hours just to make ends meet. It forms the background that led to the rise of Isolated Sign Language Recognition.

1.3 Information about PopSign's educational games

PopSign is a smartphone game app that makes learning American Sign Language fun, interactive, and accessible[8]. The primary target audience for PopSign includes parents of deaf children, though the application is readily accessible to any individual interested in acquiring ASL vocabulary. The application's learning mechanism entails users pairing ASL sign videos with corresponding bubbles containing written English words, thereby promoting the association between signs and their respective meanings.

To make ASL learning more accessible and engaging for a wider audience, real-time sign language recognition technology is considered to be added to the current PopSign game[8], which is why the Isolated Sign Language Recognition competition was proposed. By adding a sign language recognizer from this competition, PopSign players will be able to sign the type of bubble they want to shoot, providing the player with the opportunity to practice the sign themselves in addition to watching videos of other people signing[3].

1.4 Requirements for this competition

To run the machine learning model with limited latency on multiple platforms, including mobile devices, it is required for people participating in this competition to use TensorFlow Lite[11]. Also, the final model for submission should require less than 40 MB in memory and perform inference with less than 100 milliseconds of latency per video.

2 Litreature Review

Researchers conducted several research about sign language recognition and body movement predictions. However, most of the research on sign language recognition is done with single images instead of videos. Also, data for this research have strong limitations: most of their data are collected from a single signer, or they only have a small number of signs as labels.

Bauer and Heinz from the Aachen University of Technology did another research on sign language recognition[1]. Their system uses a Hidden Markov Model(HMM) to handle sequential data. They use hand shape, hand orientation, and location as feature vectors for the HMM. One thing to be noticed is that they only use the dominant hand's shape and orientation as input, and due to the research being done in 2000, they use a colored glove to help their model to determine the hand shape, which is inefficient and cannot track fingers' status. Their data set is relatively small; They only chose 97 different signs and collected videos for training and testing data from one person. This is probably why their accuracy rate ranges from 94% to 2.2%.

Fingerspelling is an important part of sign language. Without finger spelling, deaf people cannot represent their names since most sign languages are built based on representing concepts. Dahamani and Larabi built a model based on an SVM classifier to help them recognize finger patterns for sign language fingerspelling recognition for images[2]. They use geometrical features like relative area and distance to help the model get the correct hand shape. Their model's accuracy rate is impressive, ranging from 85% to 97% for three data sets. Their model's performance under complex conditions is still above 90%.

In 2022 Iyer and his colleagues did a sign language detection based on LSTM neural network model with Adam optimizer and cross-entropy as the loss function[5]. They successfully predicted dynamic signs. For the LSTM, they set up 3 sets of LSTM layers with relu activation function. They reached a high accuracy rate at 90.8% for the training data and 87.5% for the test data set. However, they did not mention their data set at all. Based on the example they provided in the article, their model only has three classes for the output.

In Kim et al.'s work *Global-Local Motion Transformer for Unsupervised Skeleton-Based Action Learning*[6], they build a transformer for unsupervised motion recognition called GL-Transformer. This transformer contains global and local attention mechanisms and a newly designed multi-task learning strategy named multi-interval pose displacement prediction, allowing GL-Transformer to simultaneously predict multiple pose displacement over different intervals. This new strategy track body's center joint and other body joints (finger, arm) separately. With their new features for GL-Transformer, they boosted the accuracy rate to above 80%.

3 Data settings

Collaborating with Google, the Georgia Institute of Technology's dataset focuses on landmarks extracted from raw videos using the MediaPipe Holistic model[7]. This extensive dataset includes 94,477 distinct body movements and 250 unique ASL signs, significantly surpassing the scope of prior research. The dataset consists of the frame number from the original video, landmark category (face, left hand, pose, right hand), landmark index, and normalized spatial coordinates[3].

To ensure the reliability and generalizability of the results obtained from analyzing this dataset, the potential bias is minimized by collecting videos from a diverse group of 21 signers from various regions across the United States. This approach aimed to capture the nuances and variations in signing styles, regional accents, and dialects present in American Sign Language. By incorporating signers with different backgrounds, the dataset becomes more representative of the broader ASL-using population, enhancing the model's ability to learn and recognize a wide range of signs[3].

It remains undecided how to combine changes in position and positions for model input. A preliminary investigation should be conducted to verify which method will work better as the input feature of the model. The sampling fidelity, which is the accuracy and quality of the landmark data extracted from samples of sign language practice videos by the MediaPipe Holistic model[7], is also waiting to be examined.

4 Proposed methods

Our solution will include three branches: training a new model under strict constraints of time and space, knowledge distillation from existing neural network-based large models, and model compression and fine-tuning from large models.

Because the input data is a sequence of landmarks, we primarily consider recurrent neural networks and transformers based architectures for the new model approach. Recurrent neural networks (RNNs) are a type of neural network that is well-suited for processing sequential data, such as time series or speech signals[9]. RNNs are designed to maintain a memory of previous inputs, which makes them ideal for modeling temporal dependencies in data. By processing the spatial coordinates of landmarks as a sequence of input signals over time, RNNs can learn to recognize different signs and sequences of signs.

Transformers are a more recent development in deep learning that has shown great promise in processing sequential data. Its architecture eliminates the need for recurrent connections, enabling highly parallelized training and significantly improving the scalability of the model.[12]. This makes them highly efficient and effective for modeling complex sequences of data. In sign language recognition, transformers could be used to learn the underlying structure of sign language gestures and recognize different signs and sequences. By attending to different parts of the input sequence at the same time, transformers could capture both the local and global context of the gestures, which could improve the overall accuracy and efficiency of the sign language recognition model.

The architectural design of this model is suggested to integrate the input feature of landmark sequences by manipulating the landmark position and changes in the positions information through designs similar to residual network [4]. Besides, processing the hand landmark data independently and combining the weight at the end of the process could potentially lead to improvement in performance as MediaPipe's hands' motion recognition allows us to use a similar method to finger spelling recognition[2], i.e., use incorporate the geometry shape of hands.

The second branch of this study will utilize knowledge distillation techniques to transfer learning from large, pre-existing neural network-based sign language recognition models to smaller, more efficient models under the teacher-student scheme. This process will involve training the smaller model to mimic the outputs of the larger model, thus retaining the valuable information and knowledge embedded within the more complex model. The performance of the distilled model will be assessed and compared to that of the original, large model to determine the efficacy of the knowledge distillation process. Both training a large model based on architectures of our new model approach and utilizing existing sign language recognition models for knowledge distillation are valuable to explore.

The final branch of this research will focus on model compression techniques, such as pruning and quantization, to reduce the size and complexity of large, pre-trained sign language recognition models without significantly impairing their accuracy or performance to meet the competition criteria. Following the compression process, these models will undergo fine-tuning to ensure high performance is retained despite the reduced size and complexity. The effectiveness of the compressed and fine-tuned models will be evaluated in terms of recognition accuracy, processing time, and computational resource requirements. It is possible to combine model compression techniques with the knowledge distillation approach to further compact models.

It is mentioned that for the competition, the model only needs to predict 5 signs simultaneously due to Popsign's game setting. Therefore, after the models are trained for 250 sign classes, a feasible approach would be to fine-tune the model to fit each individual game setting and check its ability for each 5-signs combination. It is an alternative to keep 250 signs for output classes with only the 5 target sign classes not suppressed.

To validate our model's final effectiveness, it would be suggested to test on outside datasets, including ASL practice videos that the model does not see. A possible source of such videos includes the program in American Sign Language at the University of Rochester. The program offers a range of courses for students of all levels in ASL interpretation or education and opportunities for students to engage with the deaf community. With negotiations with the program coordinator, it could be a valuable resource of ASL video datasets for model training and testing.

5 Future Work

To go beyond this, we can develop a deep learning model to translate sign language sentences into English. The challenge for a sign language sentence translator is determining which concept a motion belongs to. Also, Google mentioned that this data set is inappropriate for ASL-to-English translation or NLP study, so we need to find another data set if we want to do future studies.

References

[1] B. Bauer and H. Hienz. "Relevant features for video-based continuous sign language recognition". In: *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. 2000, pp. 440–445. DOI: 10.1109/AFGR.2000.840672.

[2] Djamila Dahmani and Slimane Larabi. "User-independent system for sign language finger spelling recognition". In: *Journal of Visual Communication and Image Representation* 25.5 (2014), pp. 1240–1250. ISSN: 1047-3203. DOI: https://doi.org/10.1016/j.jvcir.2013.12.019. URL: https://www.sciencedirect.com/science/article/pii/S1047320313002332.

[3] Google. *Google - Isolated Sign Language Recognition*. Accessed: 2023-03-31. 2023. URL: https://www.kaggle.com/competitions/asl-signs/overview.

[4] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[5] Vishwa Hariharan Iyer et al. "Sign Language Detection using Action Recognition". In: *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. 2022, pp. 1682–1685. DOI: 10.1109/ICACITE53722.2022.9823484.

[6] Boeun Kim et al. "Global-local motion transformer fornbsp;unsupervised skeleton-based action learning". In: *Lecture Notes in Computer Science* (July 2022), pp. 209–225. DOI: 10.1007/978-3-031-19772-7_13.

[7] *MediaPipe Holistic*. Accessed: 2023-03-31. URL: https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html.

[8] Tanmoy Panigrahi. *Pop Sign Learning: A bubble-shooter game that helps its players learn American Sign Language while playing*. Accessed: 2023-03-31. 2022. URL: https://devpost.com/software/pop-sign-learning.

[9] Robin M. Schmidt. *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*. 2019. arXiv: 1912.05911 [cs.LG].

[10] Hand Speak. *Signs for ABANDON*. the word Abandon in American Sign Language. 2023. URL: https://www.handspeak.com/word/3/.

[11] *TensorFlow-Lite*. Accessed: 2023-03-31. URL: https://www.tensorflow.org/lite.

[12] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].