Advancing Accessibility: ASL Visual Recognition Technology through EfficientNet

Priyanka Mazumdar[0009−0002−2564−2811],
Anindya Biswas[0009−0005−6735−4211],
Anirban Naskar[0009−0009−2570−3048],
Aritra Mandal[0009−0008−1841−5120], and
Soumya Sen[0000−0002−9178−6410]

Abstract American Sign Language (ASL), a dynamic visual-gestural language, is utilized by the Deaf community in the US, Canada, and globally. Unlike spoken languages, ASL employs handshapes, expressions, and postures for communication. Recent strides in machine learning and computer vision drive the creation of automated ASL recognition systems, bridging the divide between Deaf and hearing individuals. Our approach proposes a CNN architecture, leveraging EfficientNetV2S through transfer learning and combining recent novel techniques like data augmentation, dropout layers, and the Nadam optimizer with L2 regularization and exponential learning rate decay. The final layer utilizes softmax probabilities to map 160px × 160px RGB images to 28 classes, enhancing communication for hearing-impaired individuals. This multi-stream CNN design achieves approximately 95% accuracy, optimized for various factors including frame rate, sign execution speed, and lighting conditions. Diverse datasets and augmentation strategies counter overfitting, yielding a model with 98.64% training accuracy and 95.01% unseen validation accuracy.

Priyanka Mazumdar
A.K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India, e-mail: priyankamazumdar0621@gmail.com

Anindya Biswas
A.K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India, e-mail: anindyakbiswas5@gmail.com

Anirban Naskar
Department of Computer Science and Engineering, University of Calcutta, Kolkata, India, e-mail: a4aninass@gmail.com

Aritra Mandal
A.K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India, e-mail: aritramandal37@gmail.com

Soumya Sen
A.K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India, e-mail: iamsoumyasen@gmail.com

1 Introduction

Communication serves as the foundation for sharing thoughts, emotions, and ideas in human interactions. Yet, for the Deaf community, communication primarily relies on Sign Language, with American Sign Language (ASL) [9] being the predominant form among various global Sign Languages. Besides North America, ASL and its derivatives are used in different parts of the world including South-East Asia and Western Africa. Although the number of ASL users who use American Sign Language has never been counted by the American Census, the latest estimate of the number of ASL users in the United States is from a report for the National Census of the Deaf Population (NCDP) by Schein and Delk in 1974 [10]. Based on the 1972 survey of the NCDP, Schein and Delk provided an estimate of the ASL user population between 250,000 and 500,000 [10]. ASL is also used as a lingua franca throughout the Deaf world, widely learned as a second language [10]. Effective communication with the hearing world remains a persistent challenge for those using American Sign Language (ASL), given its distinct visual nature and unique grammar. Bridging this gap is essential to enable full participation of Deaf individuals in social, educational, and professional contexts, fostering inclusivity.

Recent advancements in machine learning [11], particularly within the realm of deep learning [16] and computer vision [4], have ignited a new era of accessibility innovation. ASL visual recognition technology powered by machine learning offers a promising solution to this challenge. By deciphering ASL gestures, translating them into text or speech, and facilitating real-time communication, this technology has the potential to redefine the landscape of communication accessibility for the Deaf community.

Our work aims to provide a CNN [2] based model to automatically detect and recognise ASL letters along with two pseudo-letters – `space` and `delete`. For this task we have used EfficientNet [20] (specifically EfficientNetV2S as available on TensorFlow [1]) which strikes the balance between size, performance, accuracy, and efficiency as compared to other similar architectures that are commonly used. This will help integrate ASL visual recognition and other successive Sign Languages into websites and communication technologies, leading to increased participation Deaf individuals in online communication, allowing other people to also converse freely with them. Furthermore, we will examine the implications of this technology for education, social interaction, and autonomous living for Deaf individuals.

By constructing an accurate and resource efficient model for ASL visual recognition, this paper contributes to broader accessibility, inclusion, and technological innovation for the Deaf community. The applications of such models extend across various domains, further augmenting their significance. In educational settings, these models can facilitate real-time translation of spoken language into sign language, fostering more effective communication between deaf students [17] and their non-deaf peers or instructors. Additionally, within workplace environments, technology-assisted solutions can empower deaf employees by providing real-time transcription

[1] https://www.tensorflow.org/api_docs/python/tf/keras/applications/efficientnet_v2/EfficientNetV2S

of verbal conversations into text or sign language, enabling seamless participation in meetings and discussions. Furthermore, the impact of deep learning models extends to healthcare and medical contexts. By leveraging CNNs for speech recognition and translation, medical professionals can communicate with deaf patients [7] more efficiently, ensuring accurate diagnosis and treatment recommendations. Similarly, in the realm of customer support, CNNs can be employed to facilitate real-time sign language interpretation during video calls, improving the overall experience for deaf customers and reducing potential miscommunication-related errors. Our trained model not only achieves better accuracy than state of the art models which relied on older architectures, but it also does so by using fewer parameters, thus shedding light on the potential of ASL recognition technology. We hope to inspire further research, collaboration, and innovation in this field, ultimately fostering a more connected and inclusive society

2 Related Work

Over the years various methods have been used to capture hand gesture data including the usage of wired gloves and depth sensors [13]. Although they provide sufficiently accurate ways of capturing gestures, such implementations are not possible everywhere because of impracticality and cost. Thus, the widespread implementation of sign language recognition relies on data capture through image sensors.

Non-automated ways of extracting relevant features from image data have been explored such as K curvature and convex hull algorithms [6]. Artificial neural networks however eliminate the need for manual feature extraction, thus enabling faster modelling and greater adaptability to variants of sign languages.

Several shortcomings are evident in prior studies [18], notably the overestimation of accuracy due to validation with inadequately sized datasets and the omission of both variation and data augmentation strategies. Such limitations hinder the generalizability of these works across diverse sign language forms and their applicability in real-world settings. Our proposed model mitigates these issues through the incorporation of an extensive dataset and the integration of data augmentation techniques. We have additionally validated our approach through a webcam-based implementation, enabling the recognition of sign language symbols displayed by participants whose images were not included in the training dataset.

Earlier image-based Sign Language Recognition works include the classification of commonly used words in Brazilian Sign Language (BSL) [1]. It achieved 81.36% accuracy on the validation set but lacked the ability to recognize alphabets. Further, it was limited to only 14 words.

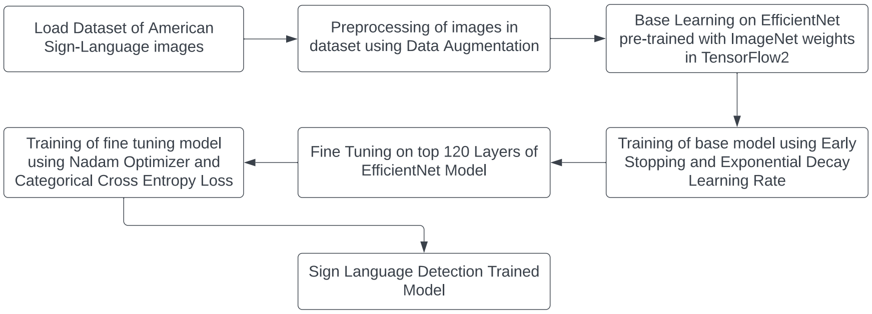
One of the previously published papers [19] employed HSV (Hue, Saturation, Value) and YCbCr formats to detect the hand symbols using skin pixels for preprocessing the image to be classified. The preprocessed images were converted into a binary format or silhouette. This silhouette dropped essential features that were required to differentiate between similar hand symbols. The most prominent examples being ‘M’ and ‘N’, where the defining feature is the position of the thumb, which cannot be distinguished from just the silhouette of the hand symbols. This introduces ambiguity during the classification of similar hand symbols.

A recent approach [15] in the domain of sign language detection for Kurdish Sign Language made use of EfficientNet for the purpose of feature extraction, while classification was only done using machine learning models. EfficientNet being a CNN trained on the ImageNet dataset, is not only good at feature extraction but also very precise at classifying symbols in real-time when fine-tuned appropriately. A machine learning model like random forest, cannot match the sophistication of a deep learning model like EfficientNet. Deep learning models are more nuanced at drawing complex patterns from high volumes of data and can optimize processing for real-time applications. This is evident from the performance of the model being higher (90% accuracy) for a small dataset on KSL and lower (84% accuracy) for a large dataset on ASL.

Another proposed work [12] used transfer learning on EfficientNet, similar to ours and reported an accuracy of 98.82%. They used a single dataset (1 of the 3 used in our work) which contained images of a single person’s hand in various lighting conditions. We noted that this dataset lacked variation in background, skin tone and hand structure. A model trained on such a dataset lacks generalization. The claimed accuracy does not translate well when a live camera feed is used for recognition [12]. The usage of data augmentation techniques in our work further increases our generalization

3 Methodology

The ImageNet dataset is a benchmark dataset that has 1000 different classes as is the output label set for ImageNet. EfficientNetV2S had its initial weights trained on the ImageNet dataset such that it performed as a general-purpose image classification model. In this model, we are using transfer learning to train the model to fit our domain which is the classification of ASL symbols.



Load Dataset of American Sign-Language images

Preprocessing of images in dataset using Data Augmentation

Base Learning on EfficientNet pre-trained with ImageNet weights in TensorFlow2

Training of base model using Early Stopping and Exponential Decay Learning Rate

Fine Tuning on top 120 Layers of EfficientNet Model

Training of fine tuning model using Nadam Optimizer and Categorical Cross Entropy Loss

Sign Language Detection Trained Model

Fig. 1 Block diagram of proposed methodology

3.1 Dataset Curation

We surveyed different datasets available on platforms such as Kaggle, GitHub, etc. and found that all the datasets for different letters of ASL were self-made by a few individuals resulting in a lack of variation in the datasets as they had multiple images with the same position of the hand, which would make the model learn spurious patterns. Hence, we decided to shuffle and merge three major ASL datasets: ASL Alphabets with a Variety of Backgrounds [14] as in Fig. 2 (b) where general household items such as windows and furniture and out-of-home scenes such as trees, buildings were set as the background to not set an unrealistic use case, ASL Alphabets with Space, Delete Symbols [5] as in Fig. 2 (c) where special symbols such as space and delete were available which we wanted to incorporate into our model's classes, ASL Alphabets with Data Augmentation [8] as in Fig. 2 (a) where a small portion of the dataset was subjected to rotation, crop and black & white filters as well as captured in front of general household items.

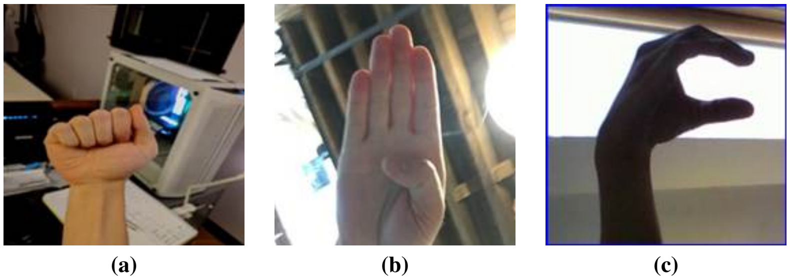


Fig. 2 Representative images picked from the datasets

The final dataset contained 3408 images which were scaled to 160px × 160px sized images before training. These belong to 28 classes, A-Z from the English Alphabet as well as `space` and `delete`.

3.2 Data Augmentation

In order to preemptively prevent overfitting on the training dataset, we have used data augmentation techniques such as applying a random 3% rotation, translation, zoom and a 50% probability of flip to introduce further variation in our chosen dataset.

3.3 Base Learning

For base learning, since the encoder part of EfficientNet has its feature extraction capability trained on the ImageNet dataset, we have used it as the base model as it is remarkably effective at embedding raw image data in the form of a vector of real numbers.

The base model (EfficientNetV2S) combined with a 20% dropout layer, followed by a fully connected dense layer with softmax activation was allowed to train on the training set for a maximum of 64 epochs.

In our training runs, early stopping stopped the base learning at the 18th epoch, which provided a training accuracy of 76.71% (loss of 0.8588) and a validation accuracy of 71.37% (loss of 1.0627). This allowed the model training to be faster and arrive at the above-mentioned accuracy without running unnecessary training epochs that do not improve the accuracy.

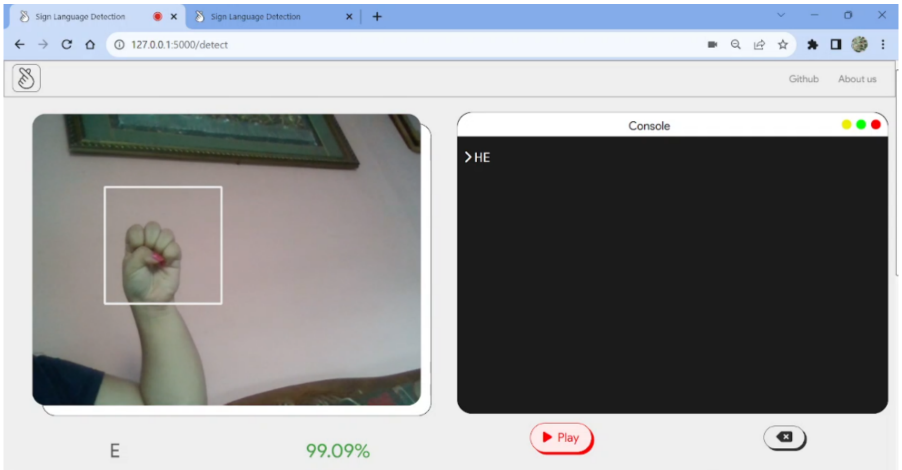
We used Nadam (Nesterov-accelerated Adaptive Momentum Estimation) Optimizer [3] with an L2 Regularizer (Ridge Regularization) for the loss and an Exponential Decay Rate as part of the deep learning techniques to reduce fitting time and lower the number of oscillations to find the weights with the lowest loss. The same learning schedule was reused for fine-tuning the model.

3.4 Fine Tuning

The EfficientNetV2S model which was trained during base learning had 35868 learn-able parameters found only in the topmost fully connected layer. However, the decoder portion of EfficientNetV2S trained on the ImageNet dataset is unsuitable for our domain, and hence we decided to further re-train the 120 top layers (20331360 learn-able parameters) of the model to fine-tune specifically for our domain of data. The fine-tuning of the model was allowed to run for a maximum of 32 additional epochs over the base learning epochs, and Early Stopping helped the fine-tuning phase conclude in 8 epochs, achieving a training accuracy of 98.64% (loss of 0.1279) and a validation accuracy of 95.01% (loss of 0.2753).

3.5 Webcam Inference

For each classification of an ASL symbol, we only considered 5 consecutive predictions of the same symbol with a confidence level of 90% in each prediction to have the predicted letter be considered for the final output. Each of these consecutive predictions was taken at 3 frames apart to further reduce the chances of misclassification. These inferred letters which form words and sentences were converted to speech for accessible communication with non-ASL speakers using WebSpeech TTS(Text-to-Speech) API.



```
>HE
```

E

99.09%

Fig. 3 Live inference from webcam capture of ASL symbol 'E'

4 Results

Our model based on the EfficientNetV2S architecture was trained on 80% of the total dataset, achieving 98.64% accuracy and a training loss of 0.1279. On the validation subset (20% of the given dataset), EfficientNetV2S retained a 95.01% accuracy with a validation loss of 0.2753, demonstrating strong generalization, as opposed to 91.78% accuracy with a validation loss of 0.3805 that was observed when trained on MobileNet. Our deployed model will only classify data when the softmax probability from the final layer exceeds 0.92.

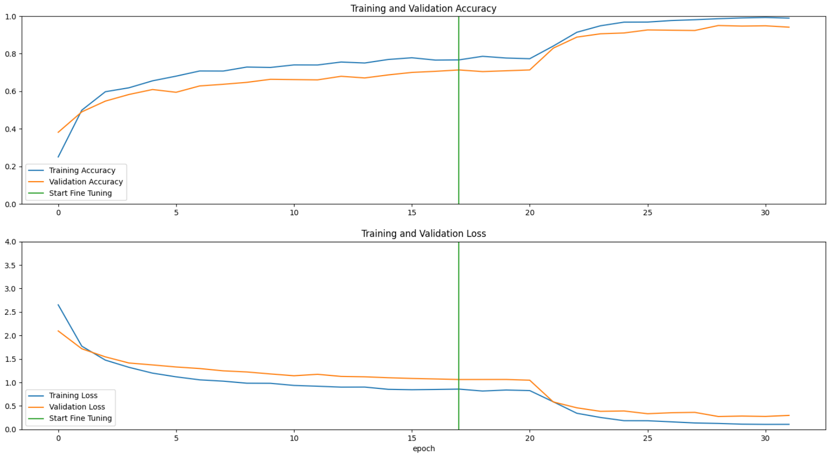


Fig. 4 Accuracy and Loss plot for training and validation sets

5 Interpretation of Results

Our version of EfficientNetV2S fine-tuned on our curated dataset performs better than the state of the art not only on the validation set but also on images of our team members' hands, which were never presented to the model during its training.

In order to explain this efficiency, we can refer to the Sparse Categorical Cross Entropy Loss plotted against epoch number Fig. 4. As is evident from the sharp decline in loss after the start of fine-tuning, the retraining of the top 120 layers is what allowed the EfficientNet architecture to adapt its decoder phase to our specific data domain. The use of Early Stopping also selected the best epoch (lowest validation loss) from 6 consecutive epochs in case the validation accuracy stopped increasing.

Adding a Dropout layer before the final Dense layer, along with L2 regularization in the Nadam optimizer also limited the amount of overfitting, while still learning enough distinct features to differentiate between the available classes. Data augmentation techniques helped improve the variation in the dataset, allowing minor rotation and flipping of images to not affect the performance of the model.

These results are in line with the performance gains noted by Tan et al. [20] which further solidified our choice of finalizing EfficientNetV2S over other model architectures for use in this case.

6 Conclusion

Traditionally, ASL interpretation has relied on human interpreters, limiting the accessibility of information for the Deaf community. Recent advancements in machine learning, computer vision, and pattern recognition have opened up new possibilities for developing automated ASL recognition systems. Our proposed solution improves over-transfer learning of pre-trained EfficientNetV2S by 33.12% through fine-tuning of the top 120 layers. The incorporated Text-to-Speech system makes communication between ASL and non-ASL speakers easier. However, due to the lack of varied datasets on commonly used words and phrases in ASL, the users of our system may find it slow in real-time communication with other users.

7 Future Scope

Letters like 'J' and 'Z' are motion-based gestures instead of static gestures. Although our CNN-based model can detect either of these with reliable accuracy, further accuracy can be achieved with the use of transformer-based architectures on a video-based dataset. That kind of transformer-based architecture would be able to harness the power of the attention mechanism to annotate relevant frames in a video for a more robust classification of motion-based gestures like in the case of 'J' and 'Z'.

This research work can also be extended, for the recognition of ASL words and phrases.

References

1. Carneiro, A.L.C., de Brito Silva, L., Salvadeo, D.H.P.: Efficient sign language recognition system and dataset creation method based on deep learning and image processing (2021)
2. Chatterjee, S., Tummala, P., Speck, O., Nürnberger, A.: Complex network for complex problems: A comparative study of cnn and complex-valued cnn. In: 2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS), pp. 1–5. IEEE (2022)
3. Dozat, T.: Incorporating nesterov momentum into adam (2016). URL https://api.semanticscholar.org/CorpusID:70293087
4. Dutta, A., Mondal, A., Dey, N., Sen, S., Moraru, L., Hassanien, A.E.: Vision tracking: A survey of the state-of-the-art. SN Computer Science (2020). DOI 10.1007/s42979-019-0059-z
5. https://www.kaggle.com/grassknoted/aslalphabet, Nagaraj, A.: Asl alphabet (2018). DOI 10.34740/KAGGLE/DSV/29550. URL https://www.kaggle.com/dsv/29550
6. Islam, M.M., Siddiqua, S., Afnan, J.: Real time hand gesture recognition using different algorithms based on american sign language. In: 2017 IEEE International Conference on Imaging, Vision and Pattern Recognition (icIVPR), pp. 1–6 (2017). DOI 10.1109/ICIVPR.2017.7890854
7. Johnson, S., Gao, G., Johnson, T., Liarokapis, M., Bellini, C.: An adaptive, affordable, open-source robotic hand for deaf and deaf-blind communication using tactile american sign language. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4732–4737 (2021). DOI 10.1109/EMBC46164.2021.9629994
8. Lee, D.: American sign language letters dataset (2020). URL https://public.roboflow.com/object-detection/american-sign-language-letters
9. Li, T., Yan, Y., Du, W.: Sign language recognition based on computer vision. In: 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 927–931. IEEE (2022)
10. MITCHELL, R.E., YOUNG, T.A., BACHELDA, B., KARCHMER, M.A.: How many people use asl in the united states?: Why estimates need updating. Sign Language Studies **6**(3), 306–335 (2006). URL http://www.jstor.org/stable/26190621
11. Mitra, A., Roy, L., Ghosh, R., Pal, D., Roy, S., Debnath, N.C., Sen, S.: Web series recommendation system using machine learning. In: Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2021, pp. 356–365. Springer (2022)
12. Patil, S., Shah, Y., Narkhede, P., Thakare, A., Pitale, R.: Gesture Detection using Tensor flow lite Efficient Net Model for Communication and E-learning Module for Mute and Deaf. International Journal of Innovative Technology and Exploring Engineering (IJITEE) **10**(8) (2021). DOI 10.35940/ijitee.H9204.0610821. URL https://doi.org/10.35940/ijitee.H9204.0610821
13. Preetham, C., Ramakrishnan, G., Kumar, S., Tamse, A., Krishnapura, N.: Hand talk-implementation of a gesture recognizing glove. In: 2013 Texas Instruments India Educators' Conference, pp. 328–331 (2013). DOI 10.1109/TIIEC.2013.65
14. Rasband, D.: Asl alphabet test from kaggle. Kaggle 2018Available online: https://www. kaggle.com/datasets/danrasband/asl-alphabet-test (accessed on 26 May 2022) (2018)
15. Salim, B., Zeebaree, S.: Kurdish sign language recognition based on efficient net. Journal of Biomechanical Science and Engineering **March 2023**, 156–170 (2023). DOI 10.17605/OSF.IO/RVS8J
16. Sarker, I.H.: Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Computer Science **2**(6), 420 (2021)
17. Senjam, S.S., Foster, A., Bascaran, C., Vashist, P., Gupta, V.: Assistive technology for students with visual disability in schools for the blind in delhi. Disability and Rehabilitation:

Assistive Technology **15**(6), 663–669 (2020). DOI 10.1080/17483107.2019.1604829. URL https://doi.org/10.1080/17483107.2019.1604829. PMID: 31012740

18. Sharma, S., Singh, S.: Isl recognition system using integrated mobile-net and transfer learning method. Expert Systems with Applications **221**, 119772 (2023). DOI https://doi.org/10.1016/j.eswa.2023.119772. URL https://www.sciencedirect.com/science/article/pii/S0957417423002737
19. Ss, S., S, D.: American sign language recognition system: An optimal approach. International Journal of Image, Graphics and Signal Processing **10** (2018). DOI 10.5815/ijigsp.2018.08.03
20. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. CoRR **abs/1905.11946** (2019). URL http://arxiv.org/abs/1905.11946