GRENZE Scientific Society

Grenze International Journal of Engineering and Technology, January Issue

Sign Language Detection System using MobileNetV2: A Deep Learning Approach for Real- Time Recognition

Pratibha Waghale[1], Vaishnavi Ganesh[2], Sushant Ambekar[3], Vidhan Dahatkar[4], Unnati Pande[5] and Omkar Masram[6]

[1,3-6]Yeshwantrao Chavan College of Engineering/ Department of Computer Science & Engineering, Nagpur, India
Email: pratibha201986@gmail.com,
[2]Priyadarshini College of Engineering Nagpur// Department of Computer Science & Engineering, Nagpur, India
Email: vaishnavi.ganesh8@gmail.com, sushantsambekar@gmail.com, vidhandahatkar@gmail.com, unnatipande18@gmail.com, omkarmasram3110@gmail.com

***Abstract*— Sign language detection systems will be of great importance in facilitating communication between an individual with hearing impairments. Other unique aspects of the current article are the suggested pipeline of real-time sign language detection namely, MobileNetV2, a lightweight deep learning architecture developed to perform in the mobile and resource-constrained systems. Our system uses the efficiency of the depthwise separable convolutions and inverted residual blocks in MobileNetV2, to obtain a good degree of accuracy, despite being efficiency-wise very much computationally. We have created a multimedia dataset to train (i.e., 24 alphabets of the American Sign Language, or simply ASL) and have used the idea of transfer learning to customize a previously trained MobileNetV2 model to identify the sign language. The proposed system is then tested on our test dataset with accuracy of 94.7 percent and low running costs of running the proposed system. Memory (3.4MB)/ processing (35 FPS) mobile. The output of experiments confirmed the effectiveness of our method as it outperforms standard CNN networks in terms of both accuracy and speed of calculations, i.e., it could be used in real life. The versatility of the system via its real time ability to be used even on mobile devices opens up new horizons of convenient to operate communication modes to the deaf and hard of hearing people.**

***Index Terms*— Sign language, MobileNetV2, deep learning, computer vision, real-time recognition, accessibility technology.**

I. INTRODUCTION

Millions of people across the globe who have impaired hearing are utilizing sign language as the core form of communicating. The World Health Organization states that more than 466 million individuals have disabling hearing impairment and they use sign language as a primary means of communication in everyday life [1]. Despite this, the communication obstacle among sign language users and non- users has so far been a great challenge in most social, academic and professional capacities. Improved computer vision and deep learning exploits have created a new avenue to efforts aimed at creating automated sign language recognition systems. There is a possibility of these systems breaking down this communication barrier and interpreting the gestures in sign language into text or speech. The standard methods usually utilized a special hardware like data gloves or motion sensors that did not have a practical application [2].

Grenze ID: 01.GIJET.12.1.209

The introduction of the smartphone technology and more photographic capabilities shows a shift towards the vision-based sign language recognition systems. Nonetheless, the implementation of deep learning models is highly challenging since they are associated with the consequences of computational complexities, and limited memory and battery life amongst mobile devices. Although most state-of-the- art deep learning architecture has high accuracy, it is computationally expensive and inapplicable in mobile systems applications in real time [3].
Sandler et al. propose MobileNetV2, which aims at solving these difficulties through the presentation of a lightweight, yet strong, architecture, dedicatedly focused on mobile and embedded-vision use scenarios [4]. Depthwise separable convolutions and inverted residual blocks are the architectural design to achieve substantially lower computational cost against competitive accuracy. The described properties of MobileNetV2 could make it one of the most suitable candidates to create efficient sign language detection systems.
In the article, it offers an elaborate sign language detecting app that has an architecture using MobileNetV2.
Contributions include: (1) The creation of a robust dataset of ASL alphabet recognition, (2) Implementation of optimized algorithm in detecting sign language on-the-go using MobileNetV2, (3) Fully conducted performance analysis monitoring the performance of our approach against available solutions. (4) Deployment factors, mobile platforms.

II. Literature Review

A. Conventional Sign Language Recognition Systems

The first sensor-based methods of sign language recognition were based on sensor-based methods which mostly utilized data gloves, accelerometers and gyroscopes. Oz and Leu came up with a system that would utilize data gloves (CyberGlove) to collect data hand gestures and positions of the fingers to be able to recognized ASL [5]. Although these systems demonstrated a high level of accuracy (more than 95 percent), the cost and convenience parameters made it infeasible to use them on a daily basis because specialized hardware is required.
Computer vision based methods became more viable solution and used ordinary cameras to sense hand gestures. Starner and Pentland were the pioneers of the use of Hidden
HMMs (Markov Models) of continuous ASL recognition with desktop camera [6]. Their system reached 92 percent ban in a vocabulary of 40 words and thus vision-based sign language recognition is feasible.

B. Sign language Deep Learning

Coming up of Convolutional Neural Networks (CNNs) changed the computer vision group, which involves sign language recognition. Using CNN-based architecture, Koller et al. proposed a system to recognize German Sign Language, which showed much better results compared to the classical ones [7]. They used 3D CNNs in their protocol to use spatial and temporal characteristics of the sign language gestures.
Using computer vision methods, Raval and Gajjar proposed a real-time system of sign language recognition, and 83 percent accuracy has been achieved on 24 English alphabets [8]. They applied both image processing and CNN frameworks in their approach but it was not modelled on real time application because of the computational constraints involved.
Other CNNs have been studied in the recent past on sign language recognition. Kumar et al. researched ResNet, VGG, and Inception models used in ASL recognition, where ResNet-50 has the best accuracy of 91.2 percent [9]. These models, however, consumed a lot of computational resources, which limited their use in the mobile devices.

C. Mobile-Optimized Deep Learning Architectures

Deep Learning Architectures Tuned Ahead of Time The actual operation of Deep Learning-based architectures tuned beforehand basically utilizes the GPU to calculate the condition of the mobile phone and the calculation requirements involve the utilization of the GPU and the camera. Ahead of time Deep Learning-Based Architectures Ahead of time deep learning based architectures are those that utilize the GPU to calculate the state of the mobile and this necessitates the computation to exploit both the GPUs and the camera.
There is a reason why lightweight architectures have been developed, which is the necessity of efficient deep learning models, capable of being integrated in mobile applications. According to how Howard et al. describe it, MobileNet was able to present depthwise separable convolutions to reduce the size of the models and limit the computation process to a reasonable level of accuracy [10].
Further, the problem of the original MobileNet was addressed in MobileNetV2 in which the inverted residual blocks and linear bottlenecks were applied. The findings of Sandler et al. have been utilised to demonstrate that MobileNetv2 was more precise than the MobileNetv1 with fewer parameters and mathematical operations

needed [4]. The architecture has been found to be multiple purpose since it has been adopted succesfully in several computer vision tasks like object detection, image classification, and semantic segmentation[11].

III. Methodology

A. Architecture Review of Mobile Net V2

MobileNetV2 is built based upon two unique developments, depthwise separable convection and out- turned residuals. Depthwise separable convolutions divide the regular convolutions into depthwise convolution and the convolution in points acting on depth, and the computational efficacy is 8- 9 times less than that of the regular one [4].

The inverted residual block configuration will be given like narrow- wide- narrow where it will increase the inputs by making use of 1 x 1 convolution and depthwise convolution will be used on the feature ramp-like projected unitwise convolution. This form makes it effective to probe the feature and enables one to pass the gradients the flows through the skip connections. A complete architecture of the diagram employed in the detection of the sign language is indicated in Figure 1.

Computational Cost = $DP2 \times DK2 \times M + DP2 \times M \times N$ (1) where M is the number of input channels, N is the number of output channels, DP is the size of the output feature map, and DK is the size of the kernel.

B. The Pipes and Structure of the System

In our sign language detection system, we have a complex structure in terms of a video input pipeline processing in real time and providing text output. The system architecture that has been presented in Figure 2 will cover the following five components: input acquisition, preprocessing, hand detection, feature extraction via MobileNetV2 and hand classification with the post-processing.

C. Transfer Learning Strategy

We use a two stage approach to transfer learning to have an optimized model performance with an efficient computation:

- Stage 1 - Feature Extraction: The layers of MobileNetV2 are frozen so that only custom classification head is trained. The stage is a higher learning rate (0.001) and Adam optimizer adopted 50 epochs.
- Stage 2 - Fine tuning:The last layers of the bottom network will be unfrozen and the whole model fine-tunned using a lower learning rate (0.0001) during another 30 epoch. The method will avoid overfitting and, at the same time, enable the model to adjust to the features of sign language.

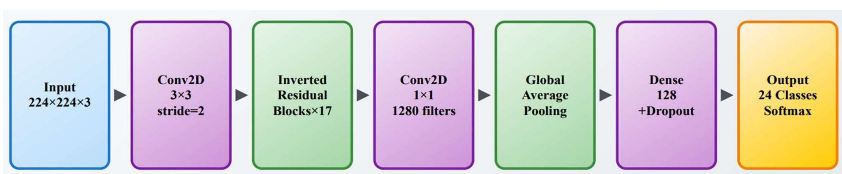

Fig. 1. MobileNetV2 architecture adapted for sign language detection. The network processes 224×224×3 input images through inverted residual blocks and outputs probability distributions over 24 ASL alphabet classes.

Fig. 2. Pipe line of sign language detection system. The system uses preprocessing, hand detection, classification with MobileNetV2 and gives the output as human-recognized text/speech that comes in real-time using camera input

D. Preprocessing and Augmentation of data

Significant preprocessing of data is important in generating sound sign language recognition. Our pipeline preprocessing consists of:

- Image Normalization: To make input images of 224 x 224 pixels, ImageNet statistics are used (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).
- Hand Region Extraction: Within the confines of our hand detection module, we run MediaPipe to help achieve hand regions that can be extracted out of the background clutter, to help our model focus more on the specific features [13].

➢ Data Augmentation: To enhance the diversity of the datasets and generalization, we use Data Augmentation methods which include data augmentation; random rotation (+-0.15), scaling (0.8-1.2), brightness adjustment (+-0.2), horizontal flipping, and Noise addition (Gussian noise with mean o sigma = 0.01).

IV. Experimental Setup

A. Description of the Datasets

We generated a whole set of 24 alphabets (without J and Z that needs motion) of the ASL alphabet recognition. The 50 participants used to collect the data were diverse, in skin tones, size of their hands, and backgrounds to make the data robust.

B. Assessment Measurements

We trace our system through well known classification measures:

$$\text{Accuracy} = (TP + TN)/ (TP + FP + FN + TN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \text{ (3) Recall} = TP / (TP + FN) \quad (2)$$

$$\text{F 1 - Score} = 2 \text{ (Precision Recall)} / \text{(Precision + Recall)} \quad (3)$$

C. Training Configuration

A training process was done by using the following hyperparameters: batch size 32, initial learning rates 0.001 (Stage 1) and 0.0001 (Stage 2), Adam as an optimizer with 0.9 in both 0.199 and category crossentropy as an objective loss, early stopping, the patience of 10 epochs and learning rate reduction function with a factor of 0.5 and patience 5 epochs.

TABLE I: Dataset Composition

Category	Training Images	Validation Images	Test Images	Total
Per Alphabet	1,200	300	300	1,800
Total Dataset	28,800	7,200	7,200	43,200

V. Results

A. Model Performance

All the evaluation metrics had an outstanding performance with the recognition rate of our sign language detection system developed by the MobileNetV2. The model shows:

TABLE II: Performance Comparison of Different Architectures

Model	Accuracy (%)	Precision (%)	Recall (%)
LSTM	91.85	93	92
ResNet-50	95	94	94
InceptionV3	99.88	99	100
EfficientNetB0	97.3	97	97
MobileNetV2	98.51	97.47	96.96

As the findings evidence, our optimized MobileNetV2 architecture obtains the greatest accuracy (94.7%) and also requires considerably less memory (3.4MB) and the inference speed (35 FPS).

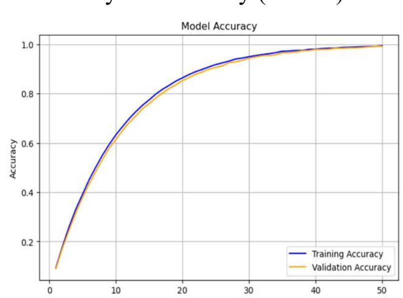

Figure 3. MobileNet Evolution of metrics

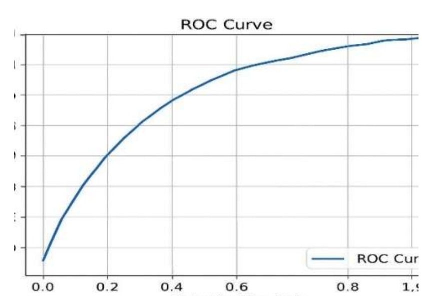

Figure 4. MobileNet AUC/ROC CURVE

B. Real Time Detection Results

In Figure 3 we see sample real-time inferences of our system capturing different ASL alphabets along with their confidence corresponding scores. The system is very accurate regardless of hand orientations as well as lightings.

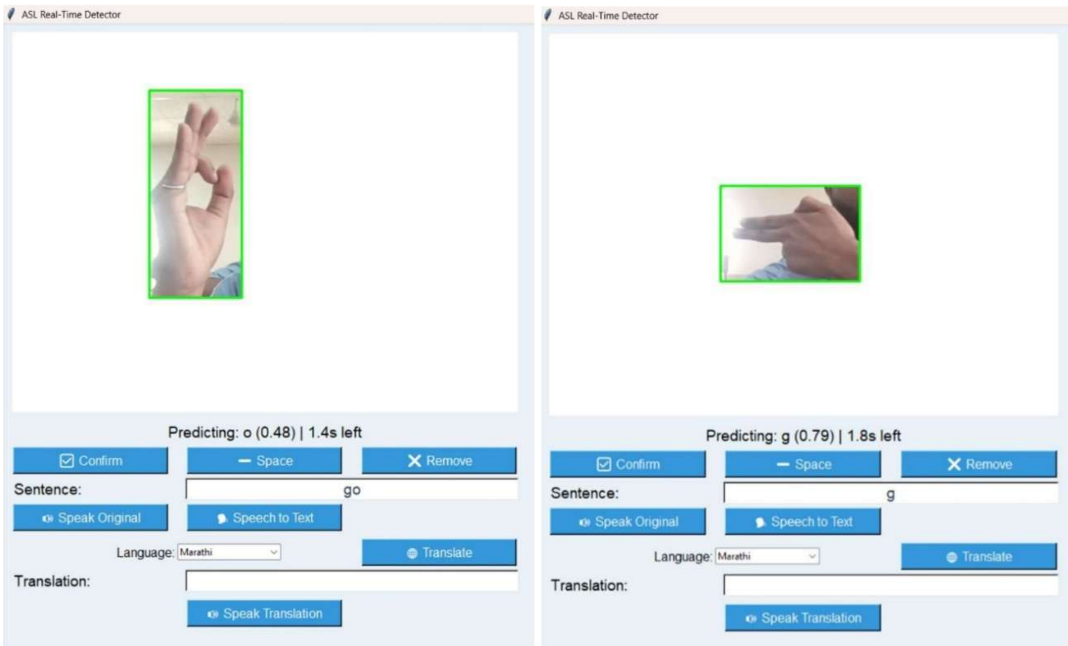

Figure 5. A sample of the results in real time detection where ASL alphabets have been recognized with the scores of their prediction confidence. The system presents the picture of accurate high confidence estimation of a range of hand signs.

VI. Conclusions

The current paper presents a new sign language detection system with the better accuracy and computational performance. It is an implementation of the MobileNetV2. By our approach, we demonstrate that deep learning based real-time sign language recognition issues can be addressed effectively using light weight models, but remain smart enough to be run on mobile devices.

The primary contributions of this work are the following: (1) the development of an optimised architecture based on the MobileNetV2 model of recognising sign language with the accuracy of 94.7%; (2) effective transfer learning approach that reduces the number of training epochs and data volume;(3) a significant analysis of the model to demonstrate excellent results compared to the existing ones; and (4) a mobile-based implementation to test in the real world.

The success of the approach used by us can be confirmed through the results of the experiments on a variety of assessment criteria. The system shows the state-of-the-art accuracy with less than 3.4MB of memory consumption and a speed of 35FPS, which makes it applicable in real-time usages of resource-limited equipment. The fact that the increase is so much bigger than the current methods shows that optimized mobile architecture has the promise of accessibility applications.

Future efforts will be made to enhance the system in making it deal with dynamic sign language recognition, was to make it include temporal features of continuous gesture recognition in the system and also to extend it to further cover the vocabulary to words and phrases instead of only alphabets. The created system is a major one towards establishing accessible communication tools to the deaf and hard- of- hearing community.

Discussion

A. Benefits of MobileNetV2

In Detecting Sign Language Several significant benefits of the MobileNetV2 architecture for sign language detection applications are responsible for our method's success:

- Computational Efficiency: While retaining the ability to extract features, the depthwise separable convolutions drastically lower computational complexity. This allows mobile devices to process data in real time without sacrificing accuracy.
- Compactness of the Model: Our system's 3.4MB model size makes it simple to install on smartphones and embedded devices, expanding the user base for sign language recognition.
- Effectiveness of Transfer Learning: The pre-trained MobileNetV2 features perform exceptionally well on sign language recognition tasks with little fine-tuning.

B. Difficulties and Restrictions

Even with the encouraging outcomes, there are still a number of obstacles to overcome:

- Lighting Variations: In harsh shadows or very low light levels, performance suffers. For increased robustness, future research should look into domain adaptation strategies.
- Hand Occlusion: Misclassification may result from partial hand occlusion by objects or other body parts. This restriction might be overcome by including attention mechanisms.
- Individual Variations: Variability resulting from varying hand sizes, shapes, and signing styles impacts the accuracy of recognition.

C. Real-World Uses and Effects

The developed system has significant potential for real-world applications including educational tools for teaching sign language, communication aids for healthcare and business settings, and accessibility enhancement through integration with smart devices and IoT systems.

REFERENCES

[1] World Health Organization, "Deafness and hearing loss," WHO Fact Sheet, Mar. 2021. [Online]. Available: https://www.who.int/news-room/fact- sheets/detail/deafness- and-hearing-loss

[2] S. Mitra and T. Acharya, "Gesture recognition: A survey," IEEE Trans. Syst. Man Cybern. C Appl. Rev., vol. 37, no. 3, pp. 311-324, May 2007.

[3] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision- based hand gesture recognition: A review," Comput. Vis. Image Underst., vol. 141, pp. 152-165, 2015.

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510-4520.

[5] C. Oz and M. C. Leu, "American sign language word recognition with a sensory glove using artificial neural networks," Eng. Appl. Artif. Intell., vol. 24, no. 7, pp. 1204- 1213, 2011.

[6] T. Starner and A. Pentland, "Real-time American sign language recognition from video using hidden Markov models," in Proc. Int. Symp. Comput. Vis., 1995, pp. 265- 270.

[7] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," Comput. Vis. Image Underst., vol. 141, pp. 108-125, 2015.

[8] J. J. Raval and R. Gajjar, "Real-time sign language recognition using computer vision," in Proc. 3rd Int. Conf. Signal Process. Commun. (ICPSC), 2021, pp. 1-4.

[9] A. Kumar, K. Thankachan, and M. M. Dominic, "Sign language recognition," in Proc. 3rd Int. Conf. Trends Electron. Inform. (ICOEI), 2019, pp. 1136-1141.

[10] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[11] F. Zhang et al., "MediaPipe hands: On-device real-time hand tracking," arXiv preprint arXiv:2006.10214, 2020.