International Journal of Research in Engineering, Science and Management
Volume 5, Issue 6, June 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

Comparative Study of CNN Based Sign Language Translation Models

Aashi Upadyay[1*], Akash Kashyap[2], Shounak Pattewale[3], Taha Bohra[4], Reetika Kerketta[5]
[1,2,3,4]Student, Department of Information Technology, MIT School of Engineering, MIT ADT University, Pune, India
[5]Professor, Department of Information Technology, MIT School of Engineering, MIT ADT University, Pune, India

***Abstract*: More than 1 million people around the world use American Sign Language. 98% of deaf individuals do not obtain sign language schooling and 70% are unemployed or underemployed. This is all due to a communication breakdown, which may be remedied with the application of technology and machine learning. This research will allow a sign language speaker and a conventional speaker to converse in a seamless manner. The task of image segmentation is to train a neural network to output a pixel-wise mask of the image. This helps in understanding the image at a much lower level, the pixel level. With the help of image segmentation, we will be training a machine to recognize a hand gesture as a sign language alphabet or word. The neural network used in the study is CNN. This deep neural network algorithm aims to solve instance segmentation problem in machine learning and computer vision. In this study, we compare the most widely used algorithm in classical Machine Learning and Deep Learning to classify the hand gestures taken from live feed input as American Sign Language. The dataset used are American Sign Language alphabets and American Sign Language basic salutation.**

***Keywords*: ASL, CNN, Hand gesture, Segmentation.**

1. Introduction

The objective of this is to build a comparative study to find out which neural network performs the best on the dataset and gives the highest accuracy to provide a sign language translation. To design an application which will create an interface to translate sign language to spoken language with the access to a learning library and to be able to translate speech to text as well as a read aloud option for impaired users.

Comparative study of different neural networks on hand gesture recognition system using sign language dataset to achieve best result for classification to translate gesture to alphabets. Taking live input of hand gestures and providing accurate results with least delay in real time translation. The algorithms will give the output in the form of classification. The algorithms used have been manipulated according to our requirements and trained on American Sign Language alphabet dataset. Since CNN works best with Image dataset; hence, it is our primary focus.

2. Related Work

A. ASL Related Research

Paper [1] demonstrates the field of HCI that pays much attention to technologies such as speech recognition and gesture recognition. In order to communicate properly with normal people, deaf and dumb individuals lack proper communication skills. An extremely high degree of accuracy is demonstrated in the project by showing how CNN can be used to solve computer vision problems. We obtain a fingerspelling sign language translator that has a 95% accuracy rate. By building the corresponding dataset and training the CNN, the project can be extended to other sign languages and different CNN models. Another paper [2] refers to the integration of SSD, Inception v3 and SVM, a vision-based American Sign Language Translator being implemented. Training of the hybrid system is implemented using the cross-validation technique and validated using a Monte Carlo estimator. The results obtained from these two experiments show the effectiveness of the cross-validation approach over the Monte Carlo method. The design of the proposed hybrid system is implemented using the ASL fingerspelling dataset. The hybrid system is its simplified structure that combines object detection, feature extraction and classification stages without ambiguities. Their designed system can only detect hands from camera images and classifies the detected object into one of the 24 signs in the ASL.

There are multiple applications that provide translation in sign language, however, there hasn't been a lightweight application developed on mobile platforms to collect live feed and translate in real-time. Our aim is to create an application that helps in achieving the goal of real-time translation from Sign Language to Text and Text to Sign Language. Since most trained models for hand gesture recognition usually are limited to translating alphabet signs into letters and then used to form sentences. However, there are a few basic applications and a few commercial web services that provide sign language translation but they either lack translating accuracy or they don't take real-time feed for translation and only work with videos and images. In order to acquire higher accuracy while keeping the application lightweight, a comparative study was drawn between MobileNet, ResNet50 and VGG16 for the ASL alphabet and numeric dataset.

*Corresponding author: ashi.08@live.com

B. CNN

We will be using convolutional neural networks to work with our dataset and classify sign language. CNN have been a recent advancement in neural networks in machine learning, despite being theorized a long time ago; we are able to use them in real-time due to advancements in hardware.

CNNs are deep neural networks, also classified as a class on neural networks nowadays. The first modern example was implemented in the form of Alexnet in 2012. It was built upon a feed forward neural network with convolution layers, the backbone of CNNs. Along with the former we also use activation layers. Together combined we get processing close to the human brain but implemented in computers.

We will be looking at the following CNNs which will be used in our comparative study:

1) VGG16

VGG16 was one of the first major improvements upon Alexnet. It was built for the Imagnet Challenge 2014, where it scored first and second place in localization and classification tracks respectively.

It's main change in the conventional CNN architecture was increasing depth of its architecture with very small 3x3 convolution filters, which showed significant improvement compared to priors works. All of this was achieved by having 16 weight layers in its architecture.

2) RESNET-50

ResNets or Residual Networks were also one of the important advances in CNNs. Resnets have an residual mapping; these use residual blocks on top of each other to form a network.

Instead of learning unreferenced functions, these use residual functions with respect to previous layers inputs. Even if this results in increased depth (152 Layers for Imagenet), the Resnet is still less complex. This results in easier optimization and considerable increase in accuracy.

3) MobileNet

MobileNet also has been one of the major advancement in CNNs, built for embedded and mobile systems. It is built upon the previous advancements we have seen in VGG and Resnet.

It uses depth-wise model like seen before, however they have introduced two global hyper parameters that tradeoff between accuracy and latency. This allows us to build light weight accurate models which are also very fast.

3. Implementation Details

We have trained all our models with Batch Size 64 on an Nvidia 3080 GPU using Imagenet pretrained weights. We have frozen the existing model layers for training and only added 3 more layers to accommodate our datasets. We have set learning rate at 0.001. And we ran each model for 50 epochs or until we reached accuracy plateau. the dataset used for training is collected from Kaggle which consists of 29 classes that includes images of English alphabets A-Z with additional three words such as delete, space and nothing(blank).

4. Results

Table 1
Results

	ResNet50	VGG16	MobileNet
mAP	0.99	0.99	0.79
Latency	76ms	60ms	64ms
Training time/epoch	780s	560s	189s

From the findings it seems like VGG16 is the best model despite it being older and heavier than MobileNet. Even though MobileNet's architecture is newer and faster the accuracy tradeoff is not worth the accuracy. In our case the models were checked ten times for the latency; so, the 64ms latency as shown in the table is not an outlier. MobileNet may compile better on embedded and mobile architectures, however VGG16 has overall better recession, accuracy, and latency on our test systems. Hence forgoing MobileNet and choosing VGG16.

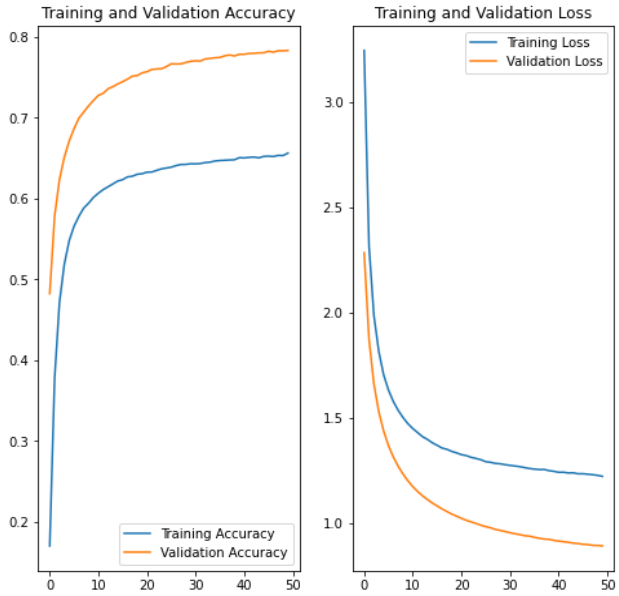

Fig. 1. MobileNet training curve

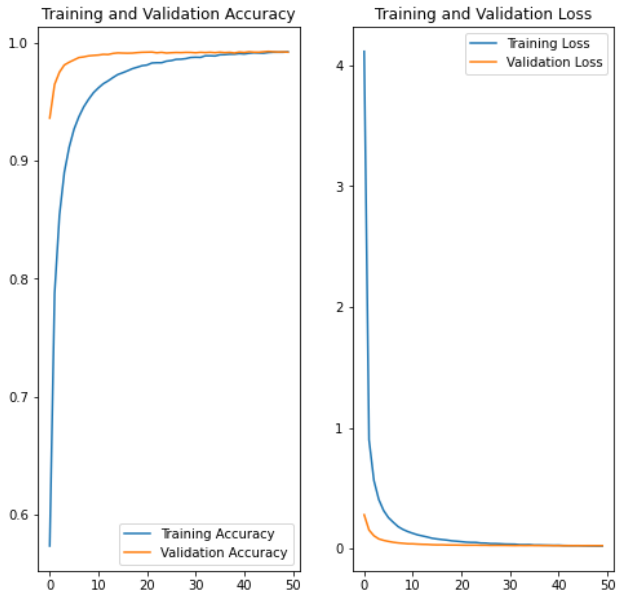

Fig. 2. VGG16 training curve

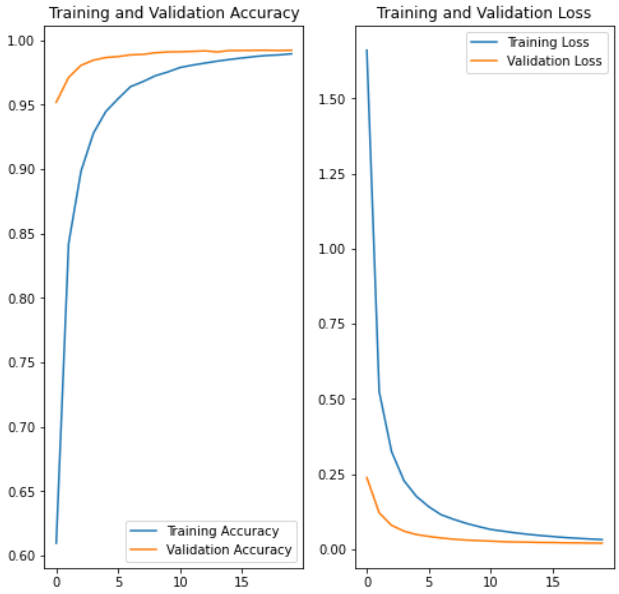

Fig. 3. Resnet training curve

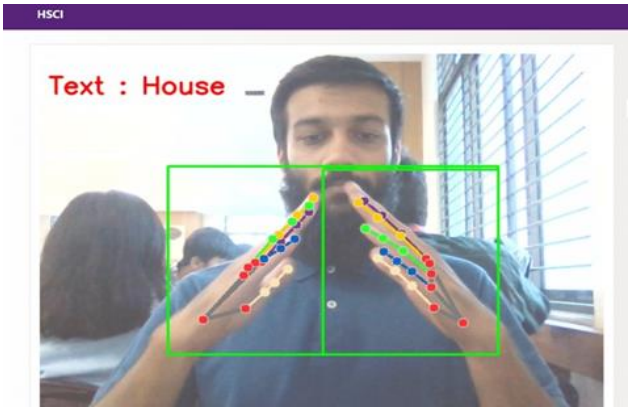

Fig. 4. Demonstration of the model working with VGG16

In reference to figure 1, 2 and 3 there is a comparison drawn between the training and validation of MobileNet, VGG16 and ResNet50 respectively. As seen in table 1, as well the figures 1, 2, 3, the time taken per epoch for MobileNet is the lowest however it takes more epoch for it to reach acceptable accuracy and loss plateau. Resnet50 and VGG16 are comparable in terms of precision and accuracy, and the time taken for them to reach accuracy and loss plateau are similar, however VGG16 achieves this in less training time, as well as being faster overall in processing images.

5. Conclusion

As seen from our results, we would conclude that VGG16 is the best architecture, as it achieves the lowest latency as well as the highest accuracy with comparative training time.

In Real-time scenarios, where we would be getting constant feedback from users in the form of new data, VGG16 would be faster to train and give us more accurate results, due to it more complex architecture; MobileNet would be only a consideration if we would using very low power embedded devices, however due to the advance in processing power in the past couple of year; MobileNet loses its novelty to more complex architectures like VGG16.

References

[1] Ankit Ojha, Ayush Pandey, Shubham Maurya, Abhishek Thakur, Dr. Dayananda P, 2020, Sign Language to Text and Speech Translation in Real Time Using Convolutional Neural Network, NCAIT – 2020, vol. 8, no. 15, 2020.

[2] R. H.Abiyev, M. Arslan and J. B. Idoko, "Sign Language Translation Using Deep Convolutional Neural Networks," KSII Transactions on Internet and Information Systems, vol. 14, no. 2, pp. 631-653, 2020.

[3] Sachin Kumar Verma, Rishabh Kesarwani, Gunjeet Kaur, "HandTalk: The interpreter for the differently abled: A Review," IJIRCT, vol. 1, no. 4, 2015.

[4] Sirshendu Hore, Sankhadeep Chatterjee, V. Santhi, and Nilanjan Dey, "Indian Sign Language Recognition Using Optimized Neural Networks," 2015 International Conference on Information Technology and Intelligent Systems (ITITS 2015), Volume: Springer-AISC.