American Sign Language Image Classification with Convolutional Neural Networks utilizing ReLU, Swish and Mish activation functions

FNU Ujjwal[1]

Abstract

The sign language is an important tool of communication for hard hearing community, and it varies from region to region. American Sign Language or ASL is prevalent for hearing or speech impaired people in North America. ASL image classification is finding its use in many mobile based applications that can help integrate deaf community better with the society. Convolutional Neural Networks or CNNs are being widely used to address this problem. This paper discusses the results of several CNN models trained on a huge ASL image classification dataset hosting 87000 images. Various experiments are conducted with CNN models by varying kernel sizes, activation functions and number of convolutional layers, to observe their effects on CNN's capability in performing ASL image classification. These trained models perform multi-class single label image classification, which classifies a particular image into one of 29 classes present. This paper also discuss how 'Swish' and 'Mish' activation function compares to 'ReLU' activation function in CNN classification models' performance. The best accuracy obtained on a non-augmented image dataset is 97.23% with (4,4) kernel size, 'Mish' activation function in a CNN having 1 Convolutional layer, 1 MaxPooling layer and 1 Dropout layer.

1. Introduction

"Language is a medium to communicate with a person or with a group. The spoken language is communication media for those who can speak and listen" (Srivastava & Malik, 2020, Pg. 1190). Language can be considered as the most significant invention of humankind and there are vast number of oral languages present in the world. But there are some people who face difficulty in hearing and speaking, and hence, are not able to utilize these languages to convey their feelings. "There are around 466 million people worldwide with hearing loss and 34 million of these are children" (Shirbhate, Shinde, Metkari, Borkar, & Khandge, 2020, Pg. 2122). These people rely on sign languages to express their thoughts.

"Sign language enables the smooth communication in the community of people with speaking and hearing difficulty (deaf and dumb)" (V. & R., 2020, Pg. 2353). Just like oral languages, a number of sign languages can also be found around the globe. "Many countries in the world have their style of Sign Language. For example, American Sign Language (ASL), Indian Sign Language (ISL), British Sign Language (BSL), French Sign Language (FSL), Chinese Sign Language (CSL), Malaysian Sign Language (MSL) and many more" (Srivastava & Malik, 2020, Pg. 1190).

Hearing and speech impaired community residing in North America mostly uses ASL for communication. ASL is also the most researched upon sign language. According to Srivastava & Malik, "American Sign Language (ASL) is a complete, complex language that uses signs made by moving the hands, facial expressions and postures of the body. It is the go-to language of many North Americans who are not able to talk and is one of the various communication alternatives used by people who are deaf or hard-of-hearing" (Srivastava & Malik, 2020, Pg. 1190). This complexity propels the need for the development of a robust and easily accessible sign recognition system "allowing anyone to understand sign languages" (Daroya, Peralta, & Naval, Jr., 2018).

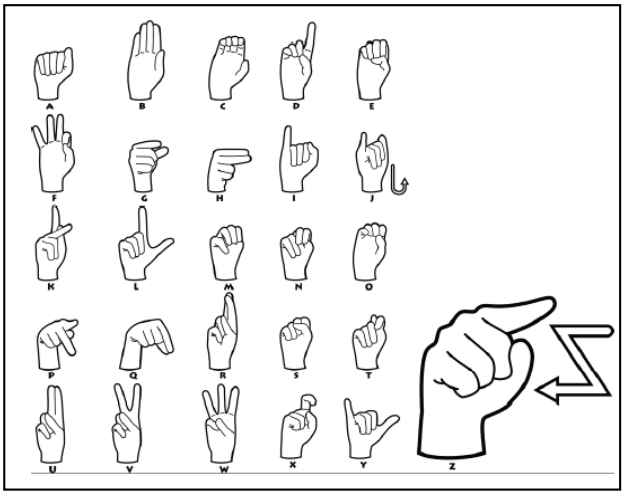


Figure 1. American Sign Language hand gestures for the alphabets (Daroya, Peralta, & Naval, Jr., 2018).

2. Problem Definition

The sign languages have played a big role in inclusivity of hard hearing community all around the world. But, very few people among the society learn these languages as it requires memorizing a lot of hand postures. This creates a communication gap between the general public and the hearing and speech impaired community. "This undeniable gap in communication is usually filled up by the help of interpreters who translates the sign language to spoken language and vice versa" (V. & R., 2020, Pg. 2353). But not everyone can afford an interpreter. Another problem faced by people who can't speak or listen is that they cannot make use of certain technologies which are making substantial advancement in today's digital era. For example, they can't utilize virtual assistants like Amazon's Alexa or Apple's Siri as they are voice controlled. Hence, development and incorporation of an efficient sign language recognition software for smart devices is eminent.

"The developments in automatic recognition of sign language gestures will be very beneficial to the deaf and dumb community as it will lead towards breaking the existing communication barrier" (V. & R., 2020, Pg. 2354). Therefore, extensive research is being carried out in the field of image classification apps that may help the general public to understand what a person using the sign language is trying to say. CNNs have been proven as an effective algorithm for image classification tasks. "CNNs automate the process of feature extraction by learning the high-level abstractions in images and capture the most discriminative feature values using hierarchical architecture" (V. & R., 2020, Pg. 2355). Therefore, this paper aims to find optimal CNN models that can confidently perform image classification task on ASL dataset. These models will be very helpful in boosting the performance of mobile applications accomplishing ASL translation and generation.

3. Related Works

A lot of research has been conducted in developing effective sign language image classification models. SVMs and hidden Markov models have also performed well in accomplishing these tasks. Naidoo et al. used SVM to classify South African Sign Language (SASL), Vogler & Metaxas used parallel hidden Markov models for ASL classification. Daroya et al. mentions that "Sole et al. used Extreme Learning Machine (ELM) to learn to classify static hand gestures on the letters of the Auslan dictionary".

Research work has also been done at the intersection of computer vision, deep learning and image processing. For example, "Tanuj Bohra et al. proposed a real-time two-way sign language communication system built using image processing, deep learning and computer vision. Techniques such as hand detection, skin color segmentation, median blur and contour detection are performed on images in the dataset for better results. CNN model trained with a large dataset for 40 classes and was able to predict 17600 test images in 14 seconds with an accuracy of 99%" (Tomar, Patel, & Thakore, 2021, Pg. 4432).

CNNs have been extensively used to perform image classification tasks. They have given impressive results for sign language image classification so far. For example, "Kshitij Bantupalli and Ying Xie worked on american sign language recognition system which works on video sequences based on CNN, LSTM and RNN. A CNN model named Inception was used to extract spatial features from frames, LSTM for longer time dependencies and RNN to extract temporal features. Various experiments were conducted with varying sample sizes and dataset consists of 100 different signs performed by 5 signers and maximum accuracy of 93% was obtained" (Tomar, Patel, & Thakore, 2021, Pg. 4432).

In another research, "Ameen et al. proposed a recognition model for letters of ASL alphabet using CNN. They utilized the features extracted from both color and depth images of gestures using two parallel CNNs and achieved a recognition accuracy of 80.34% on the ASL finger spelling benchmark dataset" (V. & R., 2020, Pg. 2355).

V. & R. mentions that "the recently emerged deep learning techniques, and advancements in convolutional neural networks (CNN) out-weighs the classical approach to hand gesture recognition as it avoids the need of deriving complex handcrafted feature descriptors from images, following the conventional pre-processing and segmentation steps" (V. & R., 2020, Pg. 2355).

4. Dataset Description

This paper uses 'ASL Alphabet – Image data set for alphabets in the American Sign Language' dataset which is uploaded by 'Akash Nagaraj' and is publicly available on Kaggle. It consists of personal images that are captured with intent to create this dataset. It is a 1 GB dataset consisting of 87,000 images of size 200 x 200 pixels packed in their respective label folder. There are a total of 29 classes and distribution of these classes is given as follows:

Table 1. Classification labels.

LABELS	INSTANCES
A-Z (26 ALPHABETS)	78,000 (3000 instances each)
SPACE	3000
DELETE	3000
NOTHING	3000

To train the model and validate the results, I use 'split-folders' library to create the train dataset (70% of the

images), validation dataset (10% of the images) and test dataset (20% of the images). The link to the dataset is as follows:

https://www.kaggle.com/grassknoted/asl-alphabet

5. Proposed Method

First, I observe the effect of varying the kernel sizes on the performance of CNN models, keeping its architecture fixed. This means that I only vary kernel size (like 4x4, 5x5, 7x7 etc.) and not any other aspect of the model (like activation functions or number of hidden layers). For these models, I use a single Convolutional layer, followed by one MaxPooling layer and one Dropout layer.

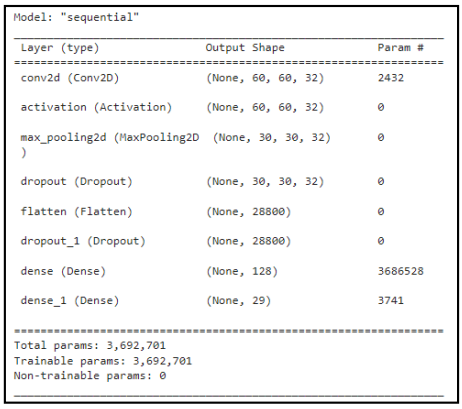


Figure 2. CNN model summary of the CNN model having (5,5) kernel size, one Convolutional layer followed by one MaxPooling layer and a Dropout layer.

Second, I observe the difference in these models' performances when I add another Convolutional layer, followed by 1 MaxPooling layer and 1 Dropout layer.

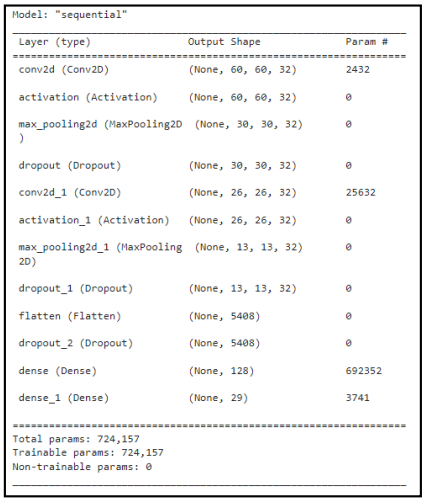


Figure 3. CNN model summary of the CNN model having (5,5) kernel size, two Convolutional layers, two MaxPooling layers and two Dropout layer.

Third, I compare the performance of all these CNN models when different activation functions are used. I explore the potential of 'Swish' and 'Mish' as an alternate to 'ReLU'.

Table 2. kernel sizes and activation functions

PARAMETERS	VALUES
KERNEL SIZES	(4,4), (5,5), (6,6), (7,7), (8,8)
ACTIVATION FUNCTIONS	ReLU, Swish, Mish

Finally, I implement data augmentation techniques, specifically, horizontal flip, vertical flip and random brightness correction, to these models and observe the difference in performances. The performance is mainly compared based on accuracy of the CNN models in classifying an ASL image.

These experiments resulted in training of 60 different CNN models and their results are discussed in the following sections of the paper.

6. Experiments & Results

6.1 Effects of varying kernel sizes

It is observed that kernel size of (4,4) gives the best result in most of the cases. For non-augmented image dataset, the best accuracy for different kernel sizes are as follows:

Table 3. best accuracy score obtained with different kernel sizes for non-augmented image dataset

KERNEL SIZES	BEST ACCURACY SCORE OBTAINED
(4,4)	97.23 %
(5,5)	96.78%
(6,6)	96.48%
(7,7)	95.33%
(8,8)	88.97%

For augmented image dataset, the best accuracy for different kernel sizes are as follows:

Table 4. best accuracy score obtained with different kernel sizes for augmented image dataset

KERNEL SIZES	BEST ACCURACY SCORE OBTAINED
(4,4)	82.44 %
(5,5)	78.98%
(6,6)	77.30%
(7,7)	74.40%
(8,8)	68.19%

6.2 Effects of adding another convolutional layer

Adding another convolutional layer doesn't improve the performance that much. In fact, for kernel sizes (6,6), (7,7) and (8,8), some of the models don't even converge. This implies that the target dimension to which the image is being converted (64x64x3), needs to be experimented with, to give these kernels more information to capture. For non-augmented image dataset, best accuracy score achieved was 97.23% and for augmented dataset, best accuracy score achieved was 82.44%.

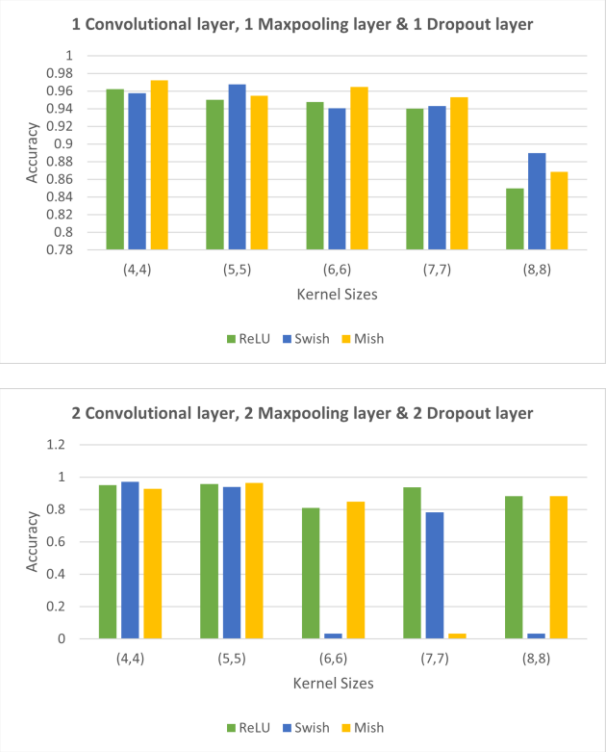


Figure 4. Prediction results with different kernel sizes and different activation functions (non-augmented image dataset)

6.3 'ReLU', 'Swish' and 'Mish'

It is observed that 'Mish' activation function comparatively performs better than 'ReLU' and 'Swish' in most of the cases.

In the following images: (1) model 1 means CNN with one Convolutional layer, one MaxPooling layer and one Dropout layer trained on non-augmented image dataset, (2) model 2 means CNN with two Convolutional layers, two MaxPooling layers and two Dropout layers trained on non-augmented image dataset, (3) model 3 means CNN with one Convolutional layer, one MaxPooling layer and one Dropout layer trained on non-augmented image dataset, and (4) model 4 means CNN with two Convolutional layers, two MaxPooling layers and two Dropout layers trained on non-augmented image dataset.

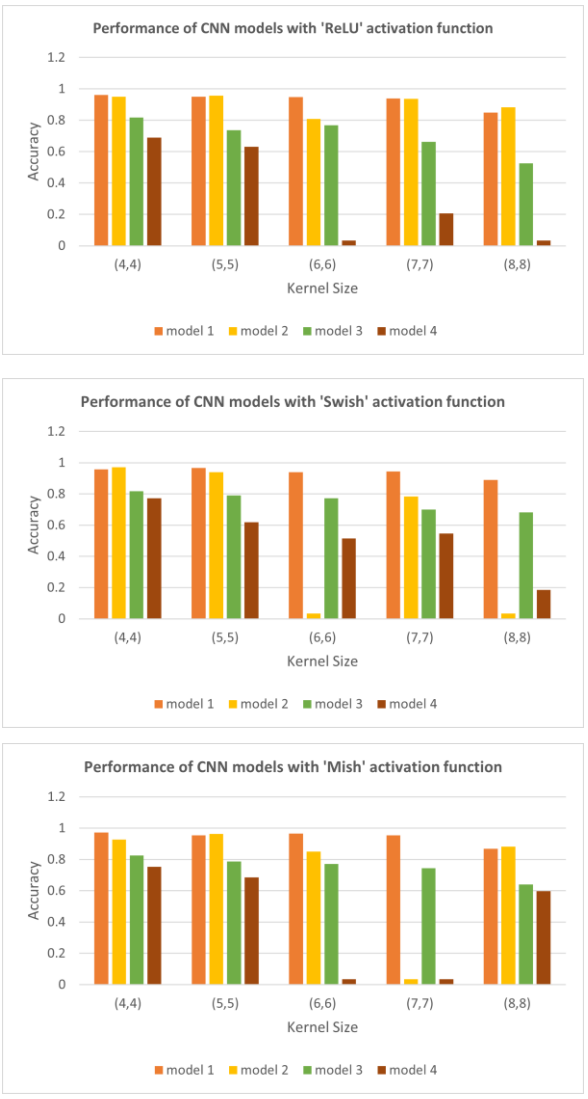
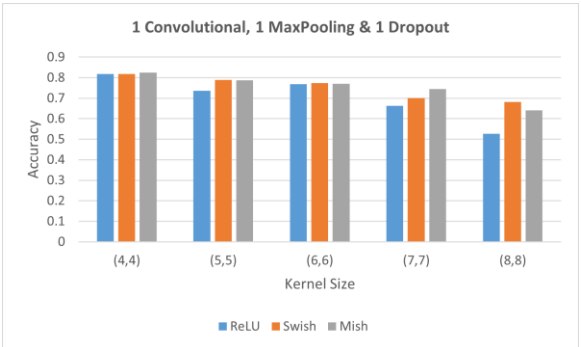


Figure 5. Performance of different CNN models with 'ReLU', 'Swish' and 'Mish' activation functions.

6.4 Effects of Data Augmentation

CNN models didn't perform well with augmented images data. For augmenting the data, random images are either horizontally flipped, vertically flipped or its brightness is varied.



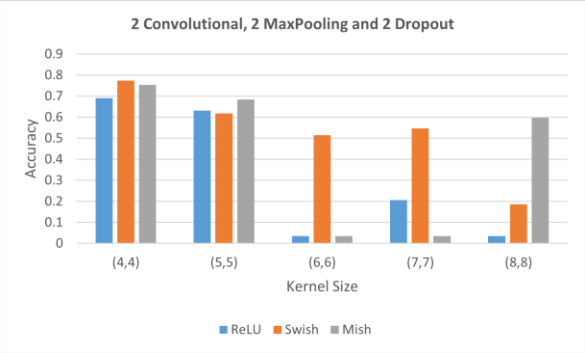


Figure 6. Prediction results with different kernel sizes and different activation functions (augmented dataset)

6.5 Best model and its training results

The best model is obtained for non-augmented image dataset, and it classifies the ASL images with 97.23% accuracy. It uses a kernel size of (4,4) and 'Mish' activation function in CNN model with 2 Convolutional layers, 2 MaxPooling layers and 2 Dropout layers.

The following figure shows how the model performed for each label:

	precision	recall	f1-score	support
A	0.93	0.99	0.96	600
B	0.96	0.99	0.97	600
C	1.00	1.00	1.00	600
D	1.00	1.00	1.00	600
E	0.99	0.89	0.94	600
F	1.00	0.99	1.00	600
G	1.00	0.96	0.98	600
H	0.96	1.00	0.98	600
I	1.00	0.99	0.99	602
J	1.00	1.00	1.00	600
K	0.99	0.99	0.99	600
L	1.00	1.00	1.00	600
M	0.97	0.98	0.97	600
N	0.95	0.97	0.96	602
O	1.00	0.96	0.98	600
P	0.99	0.98	0.99	600
Q	0.98	0.99	0.99	600
R	0.96	0.99	0.98	600
S	0.95	1.00	0.97	600
T	0.99	0.95	0.97	600
U	0.90	0.95	0.93	600
V	0.98	0.74	0.85	600
W	0.85	0.97	0.91	600
X	0.98	0.93	0.96	600
Y	0.92	0.98	0.95	600
Z	1.00	0.97	0.98	600
del	1.00	1.00	1.00	600
nothing	1.00	1.00	1.00	600
space	1.00	1.00	1.00	600
accuracy			0.97	17404
macro avg	0.97	0.97	0.97	17404
weighted avg	0.97	0.97	0.97	17404

Figure 7. Precision, recall and f1-score for each class along with macro avg and weighted avg precision, macro avg and weighted avg recall, and macro avg and weighted avg f1-score for the overall model.

From the above figure, it can be concluded that the model is most confident in predicting 'C', 'D', 'F', 'G', 'I', 'J', 'L', 'O', 'Z', 'del', 'nothing' and 'space'. It can also be concluded that it is least confident in predicting 'W'.

7. Conclusions

In this research, extensive experiments are conducted to build and train CNN models to classify ASL images. All the models are trained on 87000 images and the result is a confident model that can classify ASL images with 97.23% accuracy is obtained. Using a kernel size of (4,4) and 'Mish' activation produced impressive results. The model still lacks the ability to produce reliable results for augmented data and further CNN architecture are needed to be tried to effectively decode the augmented image dataset. Parameter selection is a crucial aspect in any model training, and it can be concluded that a lot of experiments can still be done to tweak the performance of CNN models for ASL image classification.

8. Future Scope of Research

This research work holds interesting prospects for further work. Future scope of this research work involves using the CNN models for real-time ASL sign language translation. This means that the CNN model can be given a video feed and the predictions can be made continuously. This model can be further enhanced by using some autocorrect models that corrects the translation provided by the CNN models giving rise to a whole new translation system. Furthermore, experiments can be conducted to observe the effects of varying the target dimension between (64,64,3) to (200,200,3) fed to CNN models in the start. Along with these experiments, effectiveness of big kernel size like (10,10), (15,15) can also be evaluated. Evaluation of varying the kernel size in same CNN models can also be done, For example, kernel of size (5,5) in the first convolutional layer and kernel of size (3,3) in the second convolutional layer. Lastly, model can be improved further to accommodate the hinderances due to augmented images.

References

Daroya, Rangel & Peralta, Daryl & Naval, Prospero. (2018). Alphabet Sign Language Image Classification Using Deep Learning. 0646-0650.10.1109 /TENCON.2018.8650241.

Adithya V., Rajesh R., A Deep Convolutional Neural Network Approach for Static Hand Gesture Recognition, Procedia Computer Science, Vol. 171, 2020, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.04.255.

Anamika Srivastava, Vikrant Malik (2020) Review on Sign Language Detection Using Machine Learning.

Journal of Critical Reviews, 7 (10), 1190-1194. doi:10.31838/jcr.07.10.234

Fierro, Atoany & Perez-Daniel, Karina. (2020). Siamese Convolutional Neural Network for ASL Alphabet Recognition. Computación y Sistemas. 24. 10.13053/cys-24-3-3481.

Rao, G.A., Syamala, K., Kishore, P.V., & Sastry, A.S. (2018). Deep convolutional neural networks for sign language recognition. 2018 Conference on Signal Processing And Communication Engineering Systems (SPACES), 194-197.

Bheda, Vivek & Radpour, Dianna. (2017). Using Deep Convolutional Networks for Gesture Recognition in American Sign Language.

Grandhi, Chandhini and Sean Liu. "American Sign Language Recognition using Deep Learning."

Garcia, Brandon. "Real-time American Sign Language Recognition with Convolutional Neural Networks."

Pigou, L., Dieleman, S., Kindermans, P., & Schrauwen, B. (2014). Sign Language Recognition Using Convolutional Neural Networks. ECCV Workshops.

S. Naidoo, C. Omlin, and M. Glaser, "Vision-based static hand gesture recognition using support vector machines," University of Western Cape, Bellville, 1998.

C. Vogler and D. Metaxas, "Parallel hidden markov models for american sign language recognition," in Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 1. IEEE, 1999, pp. 116–122.

M. M. Sole and M. Tsoeu, "Sign language recognition using the extreme learning machine," in AFRICON, 2011. IEEE, 2011, pp. 1–6.

T. Bohra, S. Sompura, K. Parekh and P. Raut, "Real-Time Two Way Communication System for Speech and Hearing Impaired Using Computer Vision and Deep Learning," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2019, pp. 734-739, doi: 10.1109/ICSSIT46314.2019.8987908.

K. Bantupalli and Y. Xie, "American Sign Language Recognition using Deep Learning and Computer Vision," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4896-4899, doi: 10.1109/BigData.2018.8622141.

Salem Ameen, Sunil Vadera (2016), "A convolutional neural network to classify American Sign Language finger spelling from depth and colour images." Wiley Expert Systems.