AJMAS
https://alqalam.utripoli.edu.ly/science/ eISSN 2707-7179

Original article

Performance Analysis of Sign Language Detection Using Deep Neural Networks and Computer Vision

Abha Bansal

Research Scholar M. Tech, College of Engineering, Roorkee, UTU, U.K, INDIA

ARTICLE INFO

DOI: 10.5281/zenodo.4059765

** **Abha Bansal.** Research Scholar M. Tech, College of Engineering, Roorkee, UTU, U.K, INDIA. Mobile phone: +8874549734.*
bansalabha25@gmail.com

Received: *30-07-2020*
Accepted: *30-09-2020*

Keywords: *Convolutional, Neural, Classifier.*

ABSTRACT

This paper related to the Method of Training a Deep Learning Model and how we have used it for the American Sign Language detection. We have trained a Convolution Neural Nets (CNN) using Keras and TensorFlow as a backend. There is multiple image manipulation done in between using Computer Vision like resizing, thresholding, RGB2GRAY and the most important is histogram analysis which helps to identify the difference in background and image. The main aim of this project is to track the gestures made by the hand in American Sign Language and translate it into English. The entire project has been coded in Python language for its versatility. Using our Convolutional Neural Network and Keras, we were able to obtain 97.07% accuracy.

Cite this article: *Bansal A. Performance Analysis of Sign Language Detection Using Deep Neural Networks and Computer Vision. Alq J Med App Sci. 2020;3(2):78-81.*

INTRODUCTION

Sign Language (SL) is an unmistakable method of correspondence that frequently goes understudied. The interpretation procedure among signs and a communicated in or composed language is officially called 'Translation, in this paper, we see American Sign Language (ASL), which is utilized in the USA and Canada and has various vernaculars. There are 22 hand shapes that relate to the 26 letters of the letters in order, and you can sign the 10 digits on one hand.

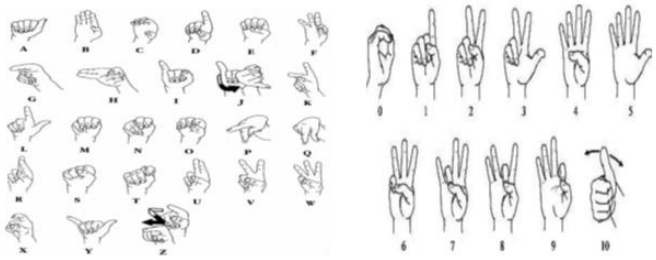

Figure1. (a) American Sign Language Alphabet (b) American Sign Language Numbers. *Data Source: https://www.kaggle.com/grassknoted/asl-alphabet*

Fingerspelling is a technique of spelling words utilizing just hand signals. One reason the fingerspelling letter set plays such an imperative role

in communication via gestures is that endorsers utilized it to explain the names of anything for which there is certifiably not a sign. Individuals' names, places, titles, brands, new nourishments, and exceptional creatures or plants all fall extensively under this class, and this rundown is in no way, shape, or form thorough. Because of this explanation, the acknowledgment procedure for every individual letter assumes a significant urgent job in its understanding.

A productive communication through signing acknowledgment framework requires information on include following and hand directions. Scientists around the globe moved toward signal order in two significant manners to be specific glove based and vision-based. Where the sensors are followed however then can't be actualized as a portable application in view of sensor location systems are as of now not present in a Mobile phone. The message will at that point be changed over from text to discourse utilizing Python's worked in help. The info will be given continuously utilizing the webcam.

Related Work

Convolutional Neural Networks have been amazingly effective in picture acknowledgment and order issues, and have been effectively executed for human signal acknowledgment as of late. Specifically, there has been work done in the domain of gesture-based communication acknowledgment utilizing profound CNNs, with input-acknowledgment that is delicate to something beyond pixels of the pictures. With the use of cameras that sense profundity and form, the procedure is made a lot simpler through creating trademark profundity and movement profiles for each communication via gestures motion [1]. The pixels of the pictures might be influenced by an alternate type[2-4]. of clamor like Brownian Noise (Fractal Noise) Rayleigh noise, gamma noise, poison-gaussian noise, salt and pepper commotion, arbitrary esteemed drive commotion, spot commotion, gaussian clamor, and structured noise, and so forth as demonstrated as follows. Commotion is incredibly hard to expel it from the mind-boggling pictures without the correct comprehension of the clamor model. The utilization of profundity detecting innovation is rapidly developing in prevalence, and different devices have been consolidated into the procedure that has demonstrated success. Developments, for example, handcrafted shading gloves have been utilized to encourage the acknowledgment procedure and make the feature extraction step progressively effective by making certain gestural units simpler to recognize and group [5-6].

DATASET AND METHODOLOGY

The CNN design we will use is a littler, progressively minimal variation of the VGGNet network [7-8]. The following is the Architecture of the Model utilized, we can see that the arrangement of convolution with max-pooling and leveling alongside the actuation layer to frame a CNN model.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 60, 60, 32)        2432
_________________________________________________________________
activation_1 (Activation)    (None, 60, 60, 32)        0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 30, 30, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 28, 28, 64)        18496
_________________________________________________________________
activation_2 (Activation)    (None, 28, 28, 64)        0
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 14, 14, 64)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 12, 12, 64)        36928
_________________________________________________________________
activation_3 (Activation)    (None, 12, 12, 64)        0
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 6, 6, 64)          0
_________________________________________________________________
flatten_1 (Flatten)          (None, 2304)              0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               295040
_________________________________________________________________
dense_2 (Dense)              (None, 29)                3741
=================================================================
Total params: 356,637
Trainable params: 356,637
Non-trainable params: 0
```

Figure2. Max pooling and Flattening along with activation layer to form a CNN model.

Let's discuss about the process in detail how a convolution network works:

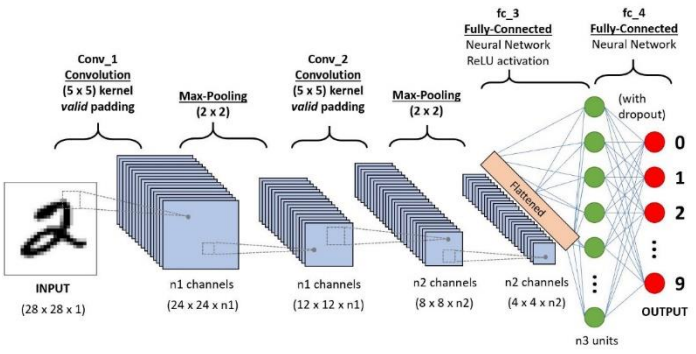

Figure 3. CNN Architecture

Above is the basic diagram for the CNN Architecture, where the input is the image to be trained. (h*w*color).

Algorithm behind Adam optimizer:

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
$m_0 \leftarrow 0$ (Initialize 1st moment vector)
$v_0 \leftarrow 0$ (Initialize 2nd moment vector)
$t \leftarrow 0$ (Initialize timestep)
while θ_t not converged **do**
$t \leftarrow t + 1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
$\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
$\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
end while
return θ_t (Resulting parameters)

Loss used as Categorical Cross-Entropy

Entropy can be calculated for a random variable with a set of x in X discrete states discrete states and their probability P(x) as follows:

H(X) = – sum x in X P(x) * log(P(x))

Below we can see the loss of model history while training the model.

As the iteration increases the Loss decreases and the Accuracy started increasing. Both the table and Graph depicts the same thing.

Table 1: Loss and Accuracy of proposed algorithms

	val_loss	val_accuracy	loss	accuracy
0	0.233053	0.923103	1.030238	0.690739
1	0.112955	0.962720	0.149178	0.950460
2	0.063774	0.979425	0.073374	0.976929
3	0.032564	0.991303	0.048675	0.984910
4	0.052743	0.983180	0.042392	0.986929
5	0.018603	0.993985	0.030158	0.991018
6	0.075645	0.978927	0.032882	0.990246
7	0.044889	0.989234	0.022527	0.993465

CONCLUSION

The paper represents the deep learning framework of Convolution Neural Nets (CNN) used with computer vision to identify the alphabets shown as a sign language. Here more focus was on prediction on running frames(video) on real-time. The audio translator is also added to the output so that the stored alphabets/words can be listened. This will help differently abled people to come up with the running future.

Future Scope

- Implementation using more robust network.
- Translation using Motion detection of finger movement
- Can be deployed using high end resource or compact model like flite or SSD that can be implemented in mobile applications,
- More data can be trained like instead of alphabets it can be trained on expressive sign languages.

Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

[1] Agarwal, Anant & Thakur, Manish. Sign Language Recognition using Microsoft Kinect. In IEEE International Conference on Contemporary Computing, 2013
[2] D. PANDEY, BINAY KUMAR PANDEY, and DR. SUBODH WARIYA, "An Approach To Text Extraction From Complex Degraded Scene", IJCBS, vol. 1, no. 2, May 2020.
[3] Digvijay Pandey., Binay Kumar Pandey, Subodh Wariya "Study of Various Techniques Used for Video Retrieval." Journal of Emerging Technologies and Innovative Research, vol. 6, pp.850-853, June 2019.
[4] Digvijay Pandey., Binay Kumar Pandey, Subodh Wariya "Study of Various Types Noise and Text Extraction Algorithms for Degraded Complex Image." Journal of Emerging Technologies and Innovative Research, vol. 6, pp.234-247, June 2019
[5] Cao Dong, Ming C. Leu and Zhaozheng Yin. American Sign Language Alphabet Recognition Using Microsoft Kinect. In IEEE International Conference on Computer.
[6] Binay Kumar Pandey, Sanjay Kumar Pandey, and Digvijay Pandey. "A Survey of Bioinformatics Applicaion on Parallel Architectures." International Journal of Computer Applications", vol. 23, pp.21-25, June 2011.
[7] Simonyan, K. and Zisserman, A. Two-stream convolutional networks for action recognition in videos. CoRR, abs/1406.2199, 2014. Published in Proc. NIPS, 2014.
[8] Pandey, D., Pandey, B.K. & Wairya, S. Hybrid deep neural network with adaptive galactic swarm optimization for text extraction from scene images. Soft Comput (2020). https://doi.org/10.1007/s00500-020-05245-4.