Sign Language Classification using Convolutional Neural Networks

Simran Popli
Western University
Ontario, Canada

Jasleen Kaur Saini
Western University
Ontario, Canada

***Abstract*—Hand gestures are an important source of communication among humans and sign language is the most effective and expressive way of communication for dumb and deaf people. These communication needs and emerging image recognition technologies aspire researchers to come up with new methods to fulfil these needs. To understand and interpret language between hearing impaired and hearing community, various approaches have been developed in the past to overcome this challenge. In this report, we present a novel method for using deep convolution networks to classify images of the letters in American Sign Language. This report provides a comprehensive study on the different approaches used for sign language recognition and also provides a framework for deep Convolutional Neural Networks (CNN), which achieves a comparable performance to the existing methods.**

***Index Terms*—Sign language recognition; machine learning; Convolutional Neural Network (CNN).**

I. Introduction

The past decade has seen a technology boom and practical applications of machine learning and artificial intelligence are yet to be realized in studying human behaviour completely. Physically disabled people, especially with hearing and speaking disabilities, can benefit from these applications. For them to communicate they use sign language which is not simple to interpret and contrary to the popular belief, sign language is not a universal language [1]. Every country has its standards of sign language which further complicates communication on a global level. In America, the American Sign Language (ASL) is considered the 3rd most used language [2]. The ASL is mainly used in the United States of America, Canada and in several countries, such as parts of West Africa and Southeast Asia [3].

According to the statistics in 2006, the number of people using ASL ranges from 250,000 to 500,000 [4]. Another research conducted for physically disabled people in India reported 2.8 Million people that depended on sign language as their primary means of communication [5]. While in Turkey similar statistical study show hearing impaired people are around 400,000 [3]. Every region has its literature and expressions because of which sign languages develop their local grammar and rules. The motivation for developing this came from the fact that it would prove to be of utmost importance for hearing impaired people and help to increase the social awareness. There are different categories of sign languages, Indian Sign Language, British Sign Language, American Sign Language etc. User can simply perform the sign in front of an image detector. The system then can then lookup the input sign with predicted results of signs stored in the system database, and presents the most similar signs (and potentially also their English translations) to the user and also produce that specific character. There have been dedicated efforts by scientist and researchers to use technology in the health-care sector and design sophisticated models for mankind. In the late 90s, the automatic sign language recognition systems were developed in different languages which made it important to develop a human-machine interaction system that can benefit people on a global level.

Fig. 1. American Sign Language Alphabet

Researchers have proposed several methods that focus on utilizing coloured gloves providing colour segmentation and extract features to distinguish the signs [4]. However, certain latest technological advancements have eliminated the use of any gloves or electric devices to develop a machine learning system. Machine learning can be defined as algorithms that can parse huge data, learn the patterns and make informed decisions based on it. Nowadays data is available in structured, unstructured form like images, text, graphics, videos as compared to traditional table-based records which call for automatic analysis methods. Since machine learning and neural networks are based on this approach of detecting, training and testing on given data; the analysis can then be utilized in detecting the label of images. This service can be used for ASL as well, since since sign language is based on hand gestures expressing the English alphabet at a time

and combining them to form a meaningful language. General image recognition systems have three fundamental phases:

1) Importing the image via image acquisition tools
2) Analysing and manipulating the image
3) Output in which result can be altered image or report that is based on image analysis

The main challenges to Sign language recognition are, feature extraction of fixed finger position, alteration due to different hand sizes, and fractional occlusion of the hand [5]. Convolutional Neural Network is the family of neural network models that feature a type of layer known as the convolutional layer which can extract features. Hence, Convolutional Neural Networks (CNN) are used for image classification since they can achieve higher accuracy. While there can be some challenges that exist in CNN for sign languages recognition like video trimming and classifying of signs. CNN's are complex feed-forward networks where multiple layers are interconnected and data is passed through such layers to improve data learning. These networks are controlled by parameters like weights and biases. We can change the parameters and build the network with a high number of iterations and a higher number of layers to yield better results, which further depends on the computational power of the system. This report aims to provide an architecture of CNN for sign language recognition, using methods which lead to a robust model with high accuracy.

II. Literature Review

Convolutional Neural Networks have been widely used in image recognition and classification problems. The dataset used for the analysis, is patterned to match closely with the classic MNIST. Simple digit and human gesture recognition systems have been successfully implemented in recent years. In particular, there have been various studies in the field of sign language recognition using deep CNNs. Therefore, the use of lenses that sense depth and contour has made the process much simpler by creating characteristic depth and motion profiles for each movement in the sign language [6]. Studies and devices such as custom-designed colour gloves have been used to facilitate the recognition process and make the feature extraction step more efficient by making gestures easy to classify and identify [8]. However, in the past image recognition based on gestures was not very successful in implementing depth-sensing technology completely. The instruments used to capture hand movements can be divided into two general groups: video-based and instrumented. The video-based approaches claim to allow the person to move freely without any instrumentation attached to the body. A camera (or an array of cameras) monitors and detects trajectory, hand shape and manual positions. All sensors are mounted on the signer's limbs or joints for instrumented approaches. The amount of data that has to be processed to extract and track hands in the image imposes a restriction on memory, speed and complexity on the computer equipment. Lately in another study, CNN's attempt to handle the classification of ASL letter gestures have been successful to a certain percentage [7]. In the application mentioned earlier [5] a system was implemented for American sign language with a finger-spelling recognition method using KNN classifier. When the pattern was denoted by full-dimensional characteristics it showed a high percentage (99.8%) for $k = 3$. The results were useful for a basic education application where the KNN classifier is more appropriate for kids to learn the ASL alphabet finger-spelling. Several previous studies implement Hidden Markov Model (HMM) as the basis of research to make Sign Language Recognition (SLR) system. HMM is a system in which the system being modelled is assumed to be a Markov process-with non-observable states. HMM has been widely deployed for the speech recognition system. HMM is also used in the glove-based Sign Language Recognition system [10]. Multi-dimensional Hidden Markov Model is used in recognizing American Sign Language (ASL) in, which has 96.7% [11]. The data from input devices (CyberGlove™) is in the form of 21 data-stream, which is then segmented into gestures in the same interval. Subsequently, the data is inputted into a 21-dimensional feature vector. The system gives an average of 95% correct recognition for the 26 alphabets and 36 basic handshapes in the ASL after it has been trained with 8 samples. The developed system forms a sound foundation for continuous recognition of ASL full signs.

Another related study [9] which is a combination of a comparison study of different language-based systems and developing a universal sign language recognition system. Where the comparison was made based on a model built based on the distance from Hausdorff and another based on the ANN. Both sign language recognition systems that were proposed in this study were used for different sign languages like American, British, and Turkish sign languages. The first system was developed by using a feed-forward neural network structure of CNN and recognition of training the dataset of each sign language. The second system was implemented based on the Hausdorff distance algorithm and Hu invariants as the focus of this approach were how to process the different movements of the hand and to recognize the different letters. OpenCV libraries were used to implement the network and it was found that the first approach gave better accuracy results of 93.4% which was ANN-based. The other gave only 90.9% success accuracy. In another study involving help by 17 volunteers with different level of skills to demonstrate 30 American Sign Language words, the system [12] attained 98% of accuracy. American Sign Language gestures are broken down into distinct phonemes sequences called Poses and Expressions, detected by software modules trained and independently tested on volunteers of different hand sizes and signing abilities. The study represented an improvement over classification based on Hidden Markov Models and Neural Network. These works and studies have motivated the continuing use of CNN to classify images and rationalize the process. Further, recent implementations in this domain also aim for a real-time machine that can automate the process.

III. Methods

We aimed to visualize and classify the images in the Sign Language MNIST dataset. The dataset contains images wherein each instance represents a single 28x28 pixel image with grayscale values between 0-255. Our overall strategy was that of simple supervised learning using stochastic mini-batch gradient descent. Our goal was to classify each letter in the dataset, using Convolutional Neural Network (CNN).

A. Data

American Sign Language Letter Database of hand gestures representing multi-class problem with 24 classes of the letter (excluding J and Z which require a motion). The training data consists of 27,455 instances and test data consists of 7172 instances which are approximately half the size of the standard MNIST. Additionally, the dataset shows no significant class imbalance as can be seen in Fig. 2.

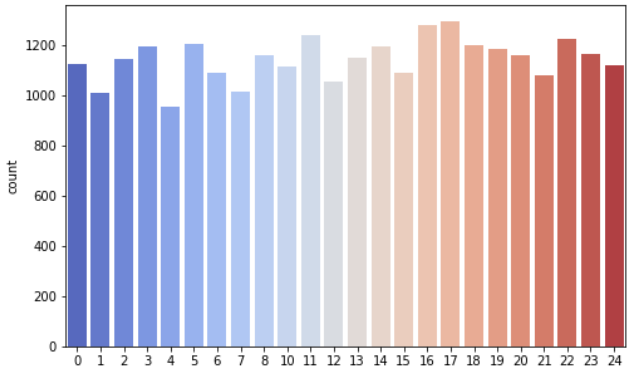

Fig. 2. Label Frequency

B. Architecture

Most of the implementations in this domain have approached it through transfer learning but our network has been trained from scratch. The architecture of our CNN model, included 2 groups of 2 convolution layers and 2 batch normalization layers followed by max-pool layer and a dropout layer, and a dense layer followed by a dropout layer and a final output layer. The network architecture is shown in Fig. 3.

In deep neural networks, the learning is based on minimizing the cost function. In the case of multi-class classification, this is the difference between the actual and predicted label. When it comes to image recognition, CNNs have the ability to detect abstract and complex features, using minimal processing. For our convolutional layers, we have used Rectilinear Unit (ReLu), since it helps to alleviate the vanishing gradient problem, which is the issue where the lower layers of the network train very slowly because the gradient decreases exponentially through the layers. Hence, by using ReLu the network can train faster. To further prevent the model from overfitting we have used a dropout layer, it's a regularization

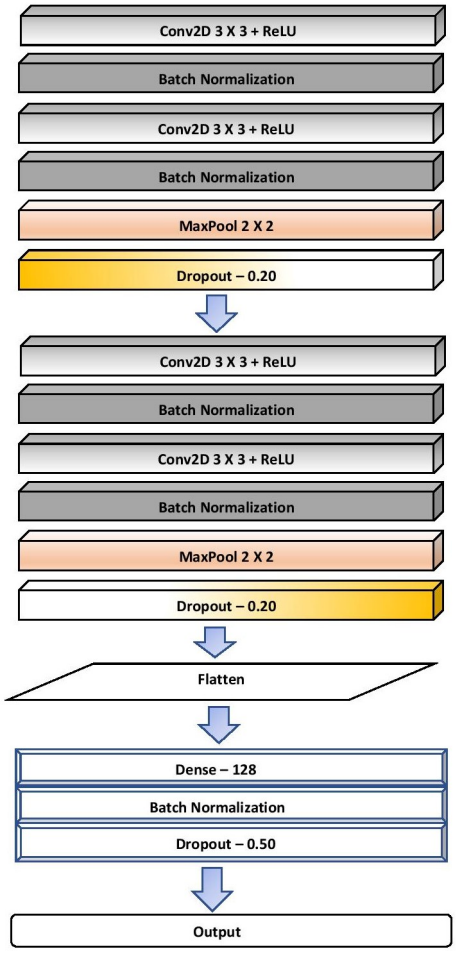

Fig. 3. Network Architecture

technique that reduces the odds of overfitting by dropping out neurons at random, during every epoch. Pooling represents another significant building block in CNNs since pooling layer decreases the size of feature maps by using some function to summarise subregions, by taking the average or maximum value which also helps in controlling overfitting.

Neural networks are susceptible to overfitting hence, to prevent overfitting in our model we have used several techniques. Firstly, we have used the Train-Validation-Test approach and we have divided our dataset into a training, validation and test set. The model is trained on the training dataset, and then the validation dataset is used to calculate the validation accuracy as to how the model is performing. Also, a test set is used to calculate the accuracy and to further analyze the model we have used a confusion matrix, which provides an insight as to how the model performs on the unseen data. Further, we also used the images of our own hands to test how the model performs on unseen data.

Additionally, to make the model more robust and less susceptible to overfitting we use data augmentation. Since in real-world applications, the images can exist in several situations like a different orientation, location, size, brightness

etc. Hence, by training our neural network with additional, synthetically changed data, we compensate for such situations. For example, we augment the data by transforming our images (rotating by 20 degrees, zoom, random shift) etc.

IV. Experimental Results

The learning rate for our model was set to 0.0001 and the optimizer used was RMSprop. RMSprop is an optimizer that utilizes the magnitude of recent gradients to normalize the gradients. RMSprop is a way to accelerate the learning process. Since our problem is a multi-class classification problem the loss function used is categorical cross entropy loss, which is commonly used for image classification problems. Also, we used a Keras neural network API with TensorFlow backend. The Keras models are built in two ways: sequential and functional. For most problems, the sequential API allows us to create models layer by layer. We used the Sequential Model API, although it does have some limitations. For example, it is not straightforward to identify models that could have: several different input sources, several destinations for outputs, or layer-reusing models.

$$H(p,q) = -\sum_{x} p(x) log(q(x))$$

Further, we split the training data and use 20% of the data as the validation set. And use a separate test data for the model to predict on, to assess the model's accuracy on unseen data. Further, we use the images of our hand to test how the model performs. The image is first pre-processed, i.e it is converted to grayscale, cropped and rescaled to 28x28 pixels. Further, we normalize the features to [0,1] before we feed them into the model for training purpose, since it reduces the effect of illumination differences and the CNN converges faster on [0,1] than on [0,255]. The neural network model is sensitive to the number of epochs and the batch size, for our analysis we tune the hyperparameters and set the number of epochs to 13 and the batch size to 64. The model performs fairly well and converges as can be inferred from Fig 4 & 5. The training accuracy is $\mathbf{99.36}\%$ and the validation accuracy is $\mathbf{99.98}\%$. To further asses the performance of our model on unseen data we use a test set. Also, we use a confusion matrix to describe the performance of the model on the test data shown in Table I. Additionally, we used images of our hands shown in Fig 6 to judge the performance of the model on new unseen images.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$f1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

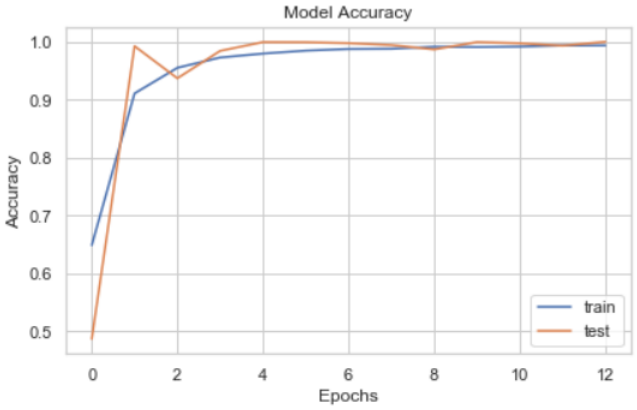

Fig. 4. Model Accuracy

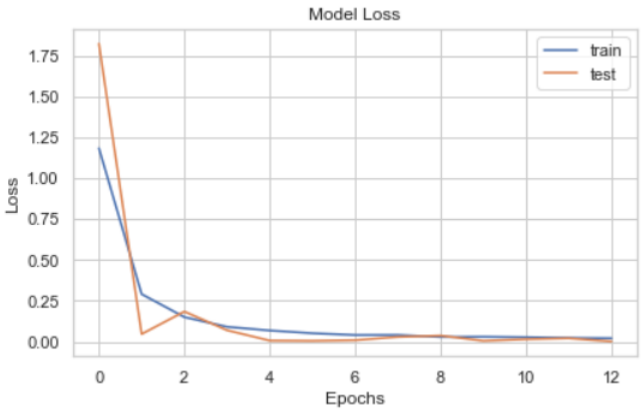

Fig. 5. Model Loss

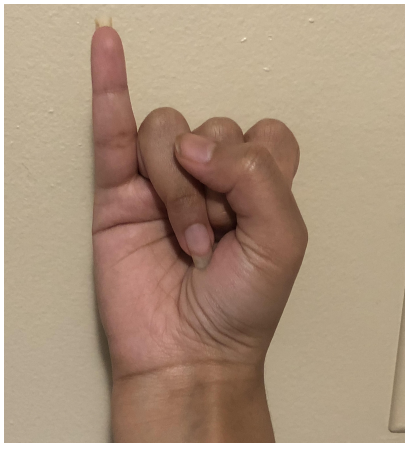

Fig. 6. Test image (Predicted Label: 8, Actual Label: 8)

label	precision	recall	f1-score	support
0	1.00	1.00	1.00	331
1	1.00	1.00	1.00	432
2	1.00	1.00	1.00	310
3	1.00	1.00	1.00	245
4	0.99	1.00	0.99	498
5	1.00	1.00	1.00	247
6	1.00	1.00	1.00	348
7	1.00	1.00	1.00	436
8	1.00	1.00	1.00	288
10	1.00	1.00	1.00	331
11	1.00	1.00	1.00	209
12	1.00	1.00	1.00	394
13	1.00	1.00	1.00	291
14	1.00	1.00	1.00	246
15	0.99	1.00	1.00	347
16	1.00	0.99	1.00	164
17	1.00	1.00	1.00	144
18	1.00	0.98	0.99	246
19	1.00	0.98	0.99	248
20	1.00	1.00	1.00	266
21	1.00	1.00	1.00	346
22	1.00	1.00	1.00	206
23	0.99	1.00	0.99	267
24	1.00	1.00	1.00	332
accuracy			1.00	7172
macro avg	1.00	0.99	1.00	7172
weighted avg	1.00	1.00	1.00	7172

TABLE I
RESULTS ON TEST DATA

V. CONCLUSIONS AND DISCUSSION

Deaf and hard hearing people use sign language to exchange information within their community and with others. Sign language detection enables automation to detect sign gestures, to produce text or speech. We can define sign movements as static and dynamic. Although, recognition of static gestures is easier than the recognition of dynamic gestures however both are important and have a large impact on our society.

Sign language recognition continues to be a significant domain for researchers to create fully automated machines to benefit the society have a larger impact. In this report, we described a convolution network approach for a classification paradigm for the American Sign Language for static sign movements. We were able to achieve good results with a high level of accuracy on both the validation and test data. The main advantage of CNN compared to its predecessors is that it automatically detects the important features without any human supervision, hence eliminating the need for traditional manual image processing methods. Hence we can further extend the architecture to extract the images from real-time videos. This will help deaf and hearing-impaired people by giving them a flexible option for interpreting when there is no face-to-face interpretation.

Hence, we can further extend the project to real-time architecture that would enable deaf people to get more involved in society. Although to extend this architecture to real-time we would need hand segmentation Since sign language is the only way for deaf people to communicate with the hearing communities as well. We can further incorporate the convolution architecture to camera-based sign language recognition system that can be used by the deaf to translate sign language gesture to text and then speech.

Also, recently there have been approaches following the zero shot learning for sign language recognition. Zero-shot learning has the ability to promote the process of translating American Sign Language into English. Implementing one-shot learning to translate the alphabet and numbers from American Sign Language to written English and compare it to a pure heuristic deep learning may be successful. Recent one-shot adaptation implementations have also been successful in solving real-world computer vision tasks, and have effectively trained deep convolutional neural networks using very little domain-specific data, including single-image data sets.

Deep learning continues to be used widely in this domain since it can extract useful features from raw data. However, it can have some limitations. Neural networks still act like a black box, and can sometimes have aberrant behaviour. Since, based on some claims of researchers deep neural networks are susceptible to adversarial instances, which are carefully selected inputs that cause the network to adjust output without making any noticeable change to a human. Besides, using deep neural networks removes the need for the Hausdorff distance that could cause some letter conflicts. Those conflicts reduce the model's ability to effectively classify characters.

REFERENCES

[1] L. Pigou, S. Dieleman, P. Kindermans and B. Schrauwen, "Sign Language Recognition using Convolutional Neural Networks", 2015.

[2] R. Mitchell, T. Young, B. Bachleda and M. Karchmer, "How Many People Use ASL in the United States? Why Estimates Need Updating", Sign Language Studies, vol. 6, no. 3, pp. 306-335, 2006.

[3] S. Ozbay and M. Safar, "Real-time sign languages recognition based on hausdorff distance, Hu invariants and neural network", 2017 International Conference on Engineering and Technology (ICET), 2017.

[4] P. Mekala, Y. Gao, J. Fan and A. Davari, "Real-time sign language recognition based on neural network architecture", 2011 IEEE 43rd Southeastern Symposium on System Theory, 2011.

[5] D. Aryanie and Y. Heryadi, "American sign language-based finger- spelling recognition using k-Nearest Neighbors classifier", 2015 3rd International Conference on Information and Communication Technology (ICoICT), 2015.

[6] Agarwal, Anant & Thakur, Manish. Sign Language Recognition using Microsoft Kinect. In IEEE International Conference on Contemporary Computing , 2013.

[7] Garcia, Brandon and Viesca, Sigberto. Real-time American Sign Language Recognition with Convolutional Neural Networks. In Convolutional Neural Networks for Visual Recognition at Stanford University, 2016.

[8] Cao Dong, Ming C. Leu and Zhaozheng Yin. American Sign Language Alphabet Recognition Using Microsoft Kinect.

In IEEE International Conference on Computer Vision and Pattern Recognition Workshops , 2015.

[9] M. SAFAR, "REAL-TIME SIGN LANGUAGES RECOGNITION BY ARTIFICIAL NEURAL NETWORKS", M. Sc. Thesis in Electrical and Electronics Engineering, UNIVERSITY OF GAZIANTEP, 2017.

[10] Liang, R. H., & Ouhyoung, M. (1998). A real-time continuous gesture recognition system for sign language. In Third IEEE International Conference on Automatic Face and Gesture Recognition (pp. 558–567). Nara, Japan.

[11] Wang, H., Leu, M. C., & Oz, C. (2006). American Sign Language Recognition Using Multi-dimensional Hidden Markov Models. Journal of Information Science and Engineering, 22(5), 1109–1123.

[12] Hernandez-Rebollar, J. L., Kyriakopoulos, N., & Lindeman, R. W. (2004). A new instrumented approach for translating American Sign Language into sound and text. In Sixth IEEE International Conference on Automatic Face and Gesture Recognition (pp. 547–552). Seoul, South Korea.