22nd INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION TECHNOLOGY

Ahsanullah University
of Science and Technology

Dept. of Computer Science and Engineering

A Novel Approach to Classify Bangla Sign Digits using Capsule Network

Tonmoy Hossain, Fairuz Shadmani Shishir, Faisal Muhammad Shah

HELLO!

I AM TONMOY HOSSAIN

Lecturer

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

INTRODUCTION

INTRODUCTION

- Hearing impaired refers to as partial or complete inability to hear
- Approximately 13 million people are suffering variable degrees of hearing loss[1]
- Previously, traditional machine learning technique was used
- Capsule Network is introduced for the classification task

[1]. Amin MN: Prevention of Deafness and Primary Ear Care (Bengali)- Society for Assistance to Hearing Impaired Children (SAHIC), Mohakhali, Dhaka-1212, Bangladesh.

MOTIVATION

- Well adaptation of automated sign digits classification in the perspective of Bangladesh
- Developing practical application for the deaf people
- On the perspective to the people who are unable to speak

RESEARCH DOMAIN

Problem

- **Classification of Bangla Sign Digits**

- ***How we can implement the problem?***
  - Traditional Machine Learning Techniques
  - Image Processing Methods
  - Deep Learning Model

BACKGROUNDS

Sign Language

- Use visual-manual modality to convey meaning
- Expressed through manual articulations in combination with non-manual elements
- Generally more than 137 types of sign language used throughout the world

Bangla Sign Digits

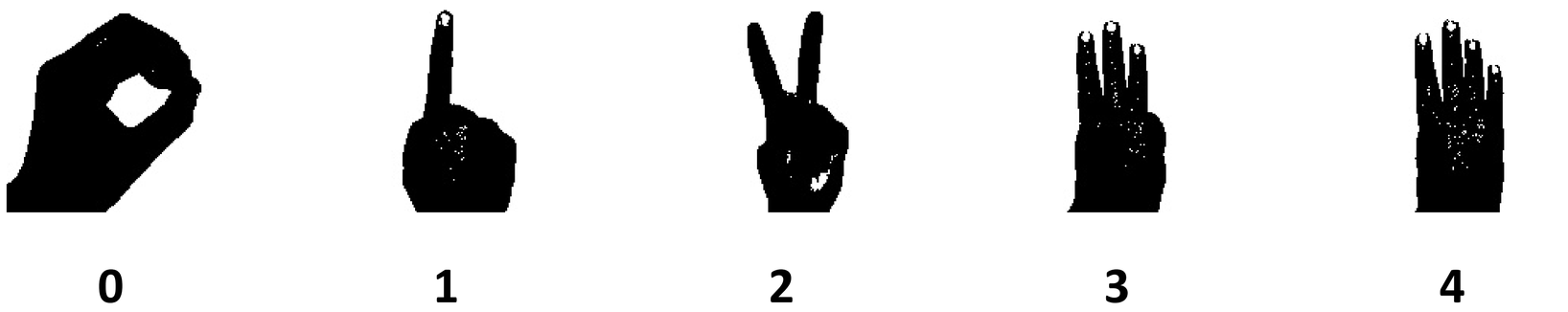


Fig 1: A depiction of Bangla Sign Digits

Capsule

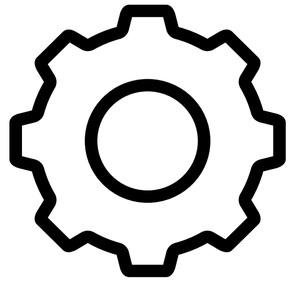
- A group of neurons (vectors) specifying the feature of the object and its likelihood
- Activity vector represents the instantiation parameter of the entity

1. Length of the activity vector represents the existence of the entity
2. Orientation to represent the instantiation parameters

BACKGROUND STUDIES

Existing Works

- **Sanzidul et al. 2018**
  - A convoluted 22 layer ConvNet architecture was implemented and achieved 94.88% of accuracy
- **Bikash et al. 2012**
  - An ensemble method of negative correlation learning and feature extraction was employed and 93% of accuracy was accomplished
- **Shahjalal et al. 2019**
  - Tracking, detecting and recognizing are the primary steps of the model which is based one data augmentation
- **Sinith et al. 2012**
  - Support Vector Machine along with Binary tree concept was operated for the classification



Proposed Methodology

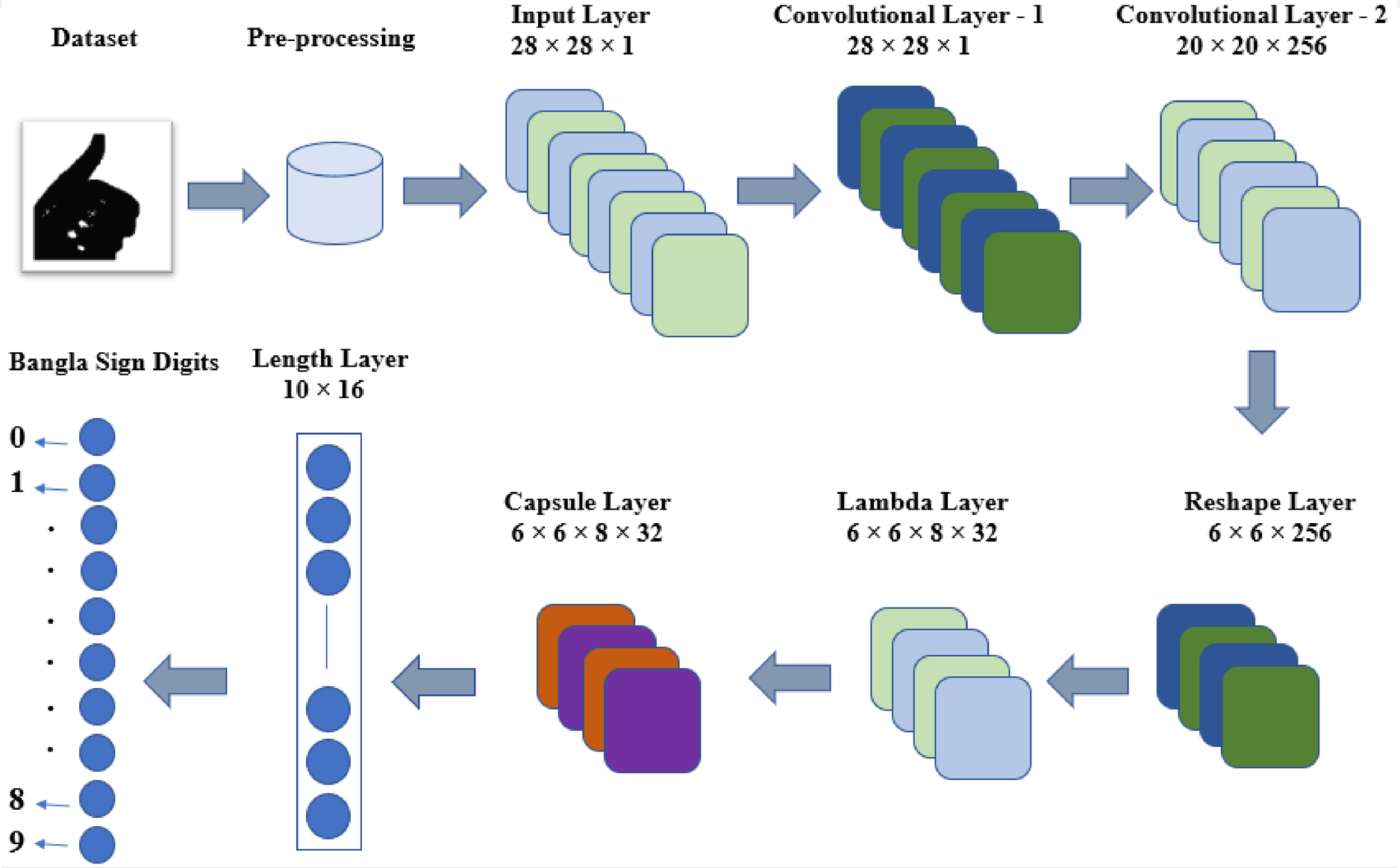


Fig 2: Proposed methodology for classification using Capsule Network

Dataset

- THE ISHARA-LIPI DATASET
- Total Images: 1000
- Break down intro ten category based on the digits
- All the images are gray scaled and binary colored
- An identical shape of 128×128 pixels is maintained in all the images

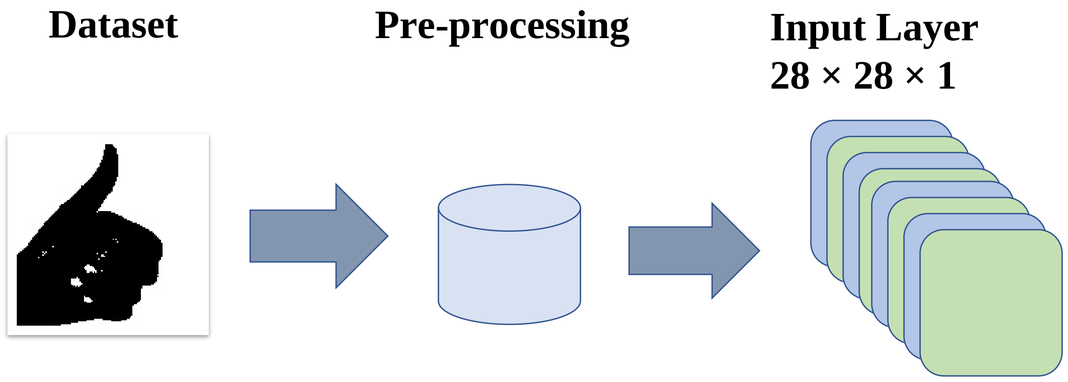
- **Some Examples**

Dataset Preparation and Pre-processing

- A total of 1000 images partitioning into 10 classes each of 100 images
- All images are converted into 28×28 pixels
- Images are labeled after binarization
- Converted the image pixels into a CSV file

Input Layer

- Pre-processed images are working as the input
- Dimension: 28×28×1
- Output a unit vector of size 10

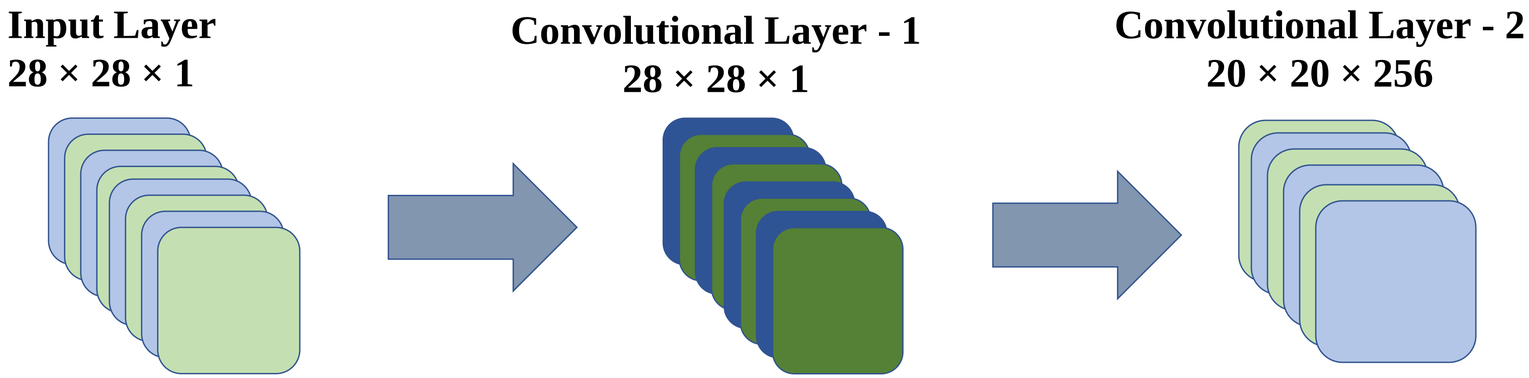


Convolutional Layer - 1

- Input Size: 28×28×1
- Output Size: 20×20×256

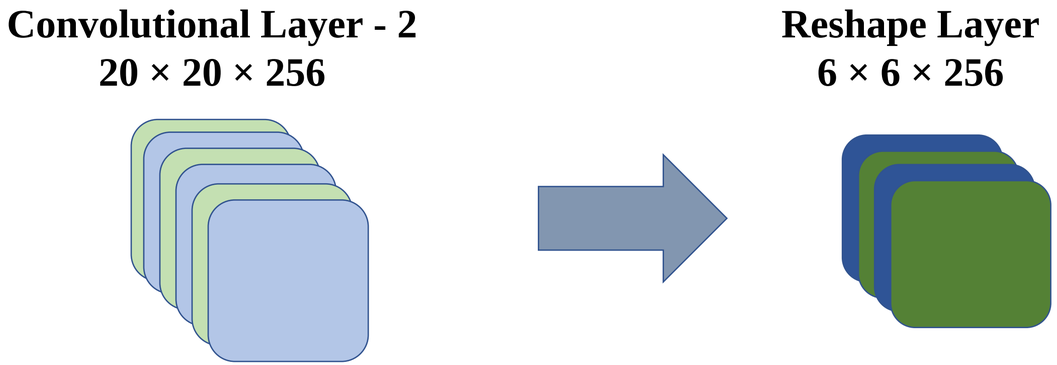
$$Output\ Size = \left(\frac{n + 2 * p - f}{2} + 1\right) \times \left(\frac{n + 2 * p - f}{2} + 1\right)$$

- Filter or Kernel Size: 9 and No padding is done
- Rectified Linear Unit (RELU) is used as the activation function
- Preserve Spatial Relation between the pixels



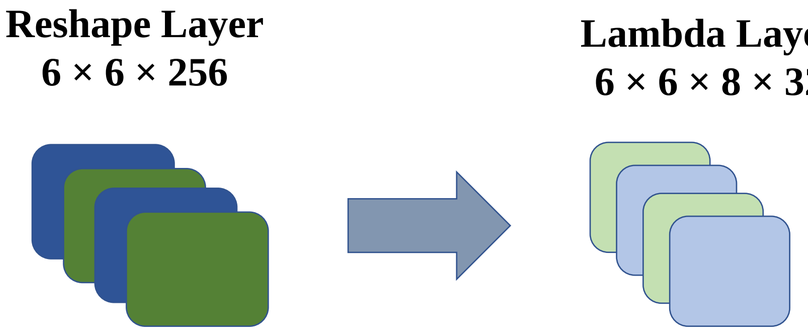
Convolutional Layer - 2

- Input Size: 20×20×256
- Output Size: 6×6×256
- Obtained a feature map after the convolutional layer



Reshape and Lambda Layer

- Primary capsule layer
- Constituted with the feature map of capsules and
- Affine transformation, weighted sum is operated
- Activation Function: Squashing Function (Non linear)



Properties of Primary Capsule Layer

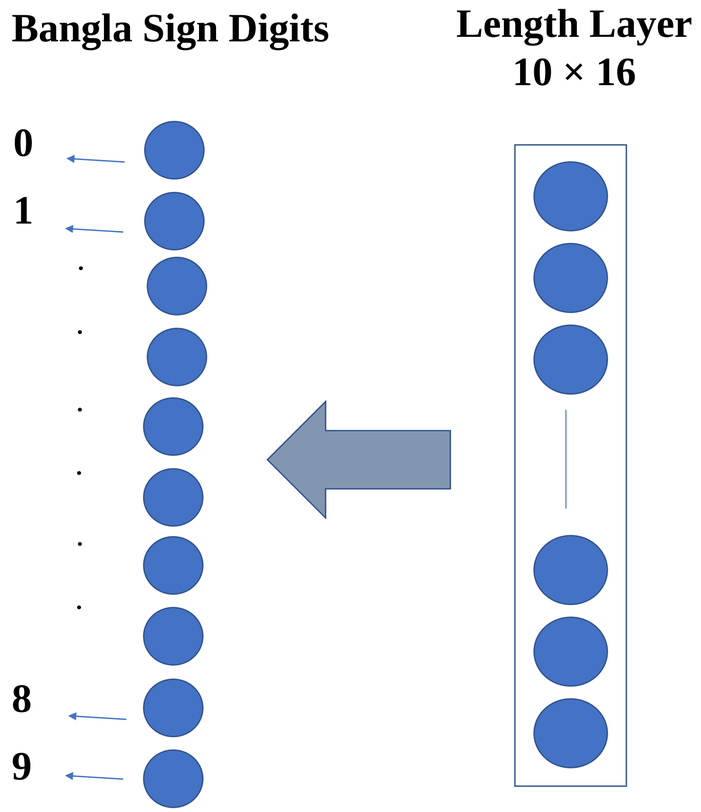
- Matrix Multiplication of input vectors with weight matrices
- Weighting input vectors
- Weighted sum
- Squashing Function

Capsule Layer (digitcaps layer)

- The higher level capsule layer
- Generate the final feature map

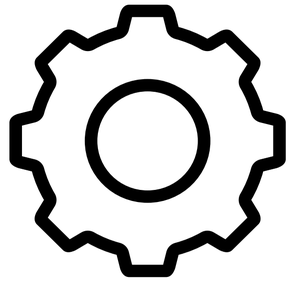
Length Layer and Classification

- Generate the final feature map
- Return the exact input shape as a tensor



A brief Workflow

A brief workflow of the proposed model	
1	Pre-processing and converted into CSV
2	Transform into an input vector
3	A feature map is obtained in the convolutional layers
4	Affine transformation is applied
5	Apply weighted sum
6	Activation function — Squashing is operated
7	A vector (shrinked) is sent to the capsule layer
8	Dynamic Routing Algorithm is performed in the capsule layer
9	Loss function is calculated for each capsule and sum up for the final loss
10	Final classified output vector is assembled



Experimental Results

Experimental Setup

- Train the data into three types of splitting ratio: 70:30, 80:20 and 90:10
- Best Result: 90:10 (80% training, 10% validation and 10% testing)
- Google Colab is used to train the model

Dimension of the Network Architecture

Layer Name	Input Shape	Output Shape	Parameters
Input Layer	(28, 28, 1)	(28, 28, 1)	0
Convolutional Layer – 1	(28, 28, 1)	(20, 20, 256)	17712
Convolutional Layer – 2	(20, 20, 256)	(6, 6, 256)	4479232
Reshape	(6, 6, 256)	(1152,8)	0
Lambda	(1152,8)	(1152,8)	0
Capsule Layer	(1152,8)	(10, 16)	1486080
Length Layer	(10, 16)	(10)	0

Table 1: Dimension of all layers of the network architecture

Hyperparameter Setup

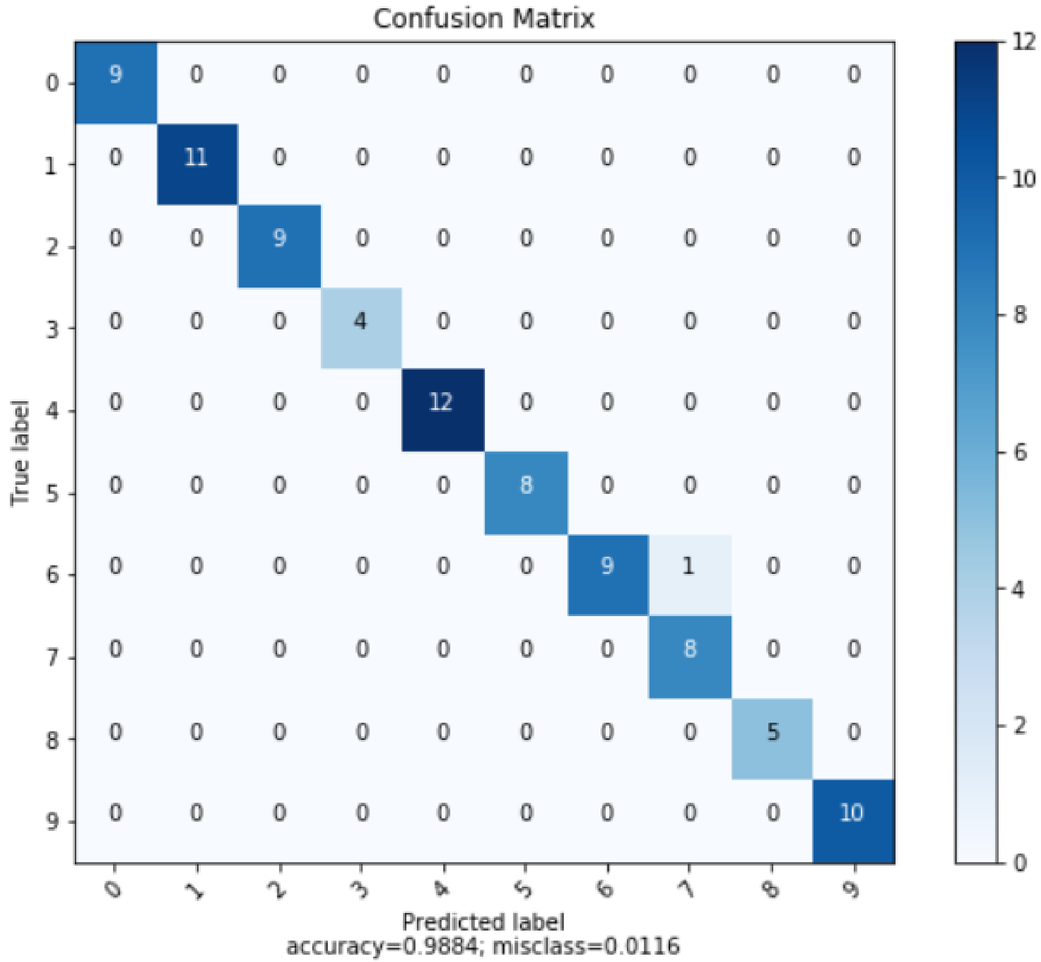
No	Hyper-parameter	Value
1	Initialization_kernel	glorot_uniform
2	Initialization_bias	zeros
3	Learning_rate	0.001
4	Optimizer	Adam
5	Batch Sie	4
6	Epoch	50
7	Steps per epoch	580

Table 2: Hyperparameters of the architecture

Classification

No	Splitting Ratio	Accuracy
1	70:30	93.92%
2	80:20	98.25%
3	**90:10**	**98.84%**

Table 3: Classification of the proposed model



Performance Comparison

Methodology	Layer	Accuracy
Sanzidul et al.	22	95.5%
Proposed Model	**7**	**98.84%**

Table 4: Performance Comparison with the existing work on same dataset

CONCLUSION

Conclusion

- The model is a seven-layer architecture which is an unsophisticated model to train and test
- Computation power and training time is curtailed with respect to the existing articles as well as data augmentation is don
- Spatial properties of an object are taking into account along with the activity vector
- Pooling layer which is lossy and does not conserve all the spatial information is apprehended in this architecture

FUTURE PLAN

Future Plan

- Work on Bangla Sign Character
- Expand the model to work on Videos
- Employ the model on a substantial large dataset
- Try to generate text from sign language conversation video

THANK YOU!