Proceedings of 12th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2022 Month: October Volume: 12

Nepali Sign Language Gesture Recognition using Deep Learning

Sanyukta Ligal [a], Daya Sagar Baral [b]

[a,b] *Department of Electronics and Computer Engineering,IOE,Pulchowk Campus,TU,Nepal*
[a] 076msdsa017.sanyukta@pcampus.edu.np , [b] dsbaral@pcampus.edu.np ,

Abstract
Sign languages play a vital role in order to develop a proper communication between the hearing impaired people and normal people. Hand gestures are most widely used sign languages which includes hand movements.In this paper we proposed a Nepali Sign Language Gesture Recognition System where both hands or single hand are used for performing gestures.Two approaches are used in which feature extraction is done by using CNN model.The classification part is done by using RNN in first approach and Vision Transformer in second approach.Results obtained from approach 2 is found to give better accuracy that approach 1.These two approaches were able to recognize word level Nepali Sign Language gestures in which percentage of 87 test accuracy has been obtained from approach 1 where RNN has been used as a classifier and percentage of 88 test accuracy has been obtained from approach 2 where Vision Transformer has been used as a classifier. VGG-16 has outperformed than two other models for features extraction part. Thus, the features extracted from VGG-16 has found to be better has compared to other two CNN models . Likewise, the results obtained shows that Vision Transformer can outperform RNN not only in machine translation task but they can also do better in task of computer vision.

Keywords
Convolution Neural Network,Vision Transformer,Gesture,Nepali Sign Language

1. Introduction

Sign languages are those languages that are mainly used for communication between the hearing impaired people and the normal people.There are many sign languages like American Sign Language, Indian Sign Language,Chinese Sign Language, etc.All the signs used in these sign languages are not similar. There is a difference in the sign languages and Nepali sign languages which are used by the deaf people are also different as compared to other sign languages.A nepali sign language teacher, Mr.Hari Prasad Adhikari has been teaching to differently abled people nepali sign language for 5 years. He has also been contributing more in virtual platform.He has been uploading different nepali sign language videos in his own youtube channel in order to make the hearing impaired people learn more about nepali sign language.There are different organizations and schools which have been helping deaf people to learn nepali sign language in our country Nepal.

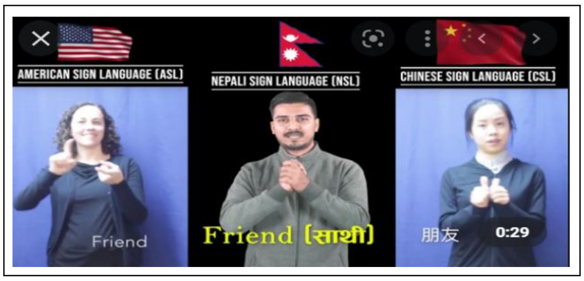

AMERICAN SIGN LANGUAGE (ASL)
NEPALI SIGN LANGUAGE (NSL)
CHINESE SIGN LANGUAGE (CSL)
Friend
Friend [साथी]
朋友
0:29

Figure 1: Different signs used for friend in ASL, NSL, CSL

There have been many sign language recognition tasks carried out for ASL, ISL,CSL,etc. But very few works have been done for Nepali Sign Language gesture recognition ,especially as per the study, tasks for Nepali Sign Language gesture recognition in word level have not been performed yet and recognition of Nepali signs for consonant and vowels are carried out.There is a communication problem between the hearing disabled people and regular people. The paper presents hybrid model for recognition of Nepali Sign Language Gestures in word level in which different hand movements are included.Previously video classification tasks are carried out using RNN and due to the presence of high trainable parameter,model is very complex.Here,classical CNN extracts the features from the images and classification task is carried out using two different approaches.RNN is used for classification in approach 1 and Vision Transformer is used for classification in approach 2. The performance of the proposed hybrid models are compared and analysed using two approaches and best model is used on the basis of high accuracy.

2. Literature Review

There have been many sign language recognition techniques being developed which can be for American Sign Language, Indian Sign Language, Chinese Sign Language, etc. But it is still a challenging research field for researchers. Maximum of works have been performed for isolated sign language recognition and less works have been done for Nepali sign language gesture recognition.

[1] developed a system for finger spelling recognition for NSL. They used the concept of vertex chain code and freeman chain code for the extraction of features of NSL.Three thinning binary images were used to test and validate the algorithms: L-block, hexagon, and pentagon.[2] developed a NSL recognition system for consonant alphabet and vowel, skin color model was used in order to segment hand from image and blob analysis was performed in order to extract hand features

A comparative analysis was proposed on Indian Sign Language (ISL)[3] gesture recognition in real time. Two different approaches were used: Euclidean distance and KNN metrics. For the feature extraction technique, this paper used direction histogram due to its appeal for illumination and orientation invariance.

[4] presented a system by using a glove based approach to recognize continuous gestures in real time. They developed a wide glossary of sign language interpreters that was useful for Taiwanese for this purpose and used time-varying parameter detection to tackle the issue of extraction for key frames. statistical analysis. They employed a Hidden Markov Model to recognize gestures. This system's average accuracy rate is 80.4. PCA for sign gesture was proposed by [5] as a quick and effective technique. They use video to extract three fps and analyze them for those gestures which are static in nature. This approach has a overall accuracy of 90.

For Korean Sign Language recognition, [6] presented a method that employs Fuzzy Logic and a HMM.Accuracy of 94 was obtained for fifteen sentences belonging to KSL using these strategies.Speed and velocity was considered for hand motion. Sarfaraz Masood [9] proposed a system to recognize hand gestures from video sequence using the hybrid model CNN and RNN which was performed for Argentinian Sign Language and produced accuracy of 95.27.Very few tasks are carried out using Vision Transformers for feature extraction and image classification.[7] trained the ViTs with ImageNet dataset and it was able to produce better accuracy as compared to CNN model.[8] proposed a system to recognize characters of ASL which includes alphabets from a to b, which used CNN based model which was able to produce accuracy of 96.

3. Methodology

A Nepali Sign Language Gesture Recognition system is proposed to be developed which means that it consists of two features which are the spatial features and the temporal features. The spatial characteristics consist of the spatial information about the individual frames of the image. It helps us to know which gestures/image is represented at that time instance.Temporal features consist of the information that takes place over a series of time. Two approaches are used for training the model for spatial and temporal features.The two approaches can be described as below.

3.1 Collection of dataset

There is no dataset available for nepali sign language on the internet due to which the dataset required for this system is created by own. The creation of the dataset required a huge amount of time and it was the most challenging task in order to create the dataset.

3.1.1 Creation of nepali sign language gesture video set

The dataset consists of a collection of signs where a single hand or both hands has been used for performing the nepali sign language gestures .In order to create dataset, we took a reference from [9] and with the approach followed in it , the dataset has been created where the individuals wore black dress and orange colored gloves. There are a total 25 signs used in this system. Thus,a total of 25 signs database has been created. The dataset consists of 25 labels arranged in a folder format in which the respective labels are represented by the folder name and all the videos for particular gestures are present inside the gesture folder. The gestures are dynamic as well as static ones. All 25 different sign language gestures are recorded. 22 non-expert subjects were used for creating the gesture videos. Each of the non-expert individuals performed the repetitions of the 25 gestures for 5 times each . The first two non-expert individuals performed 10 repetitions of the 25 gestures and the remaining 20 individuals performed 5 repetitions. We took help from more individuals in order to have variation in the dataset . There are altogether ten two handed signs and 15 single handed signs.

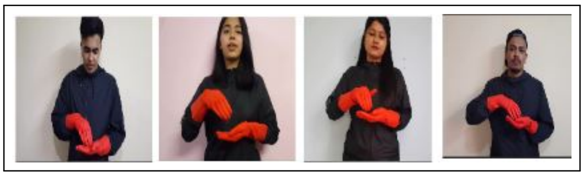


Figure 2: 4 Sample images of different individuals for gesture ‘Chaamal’

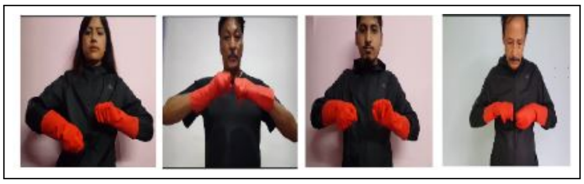
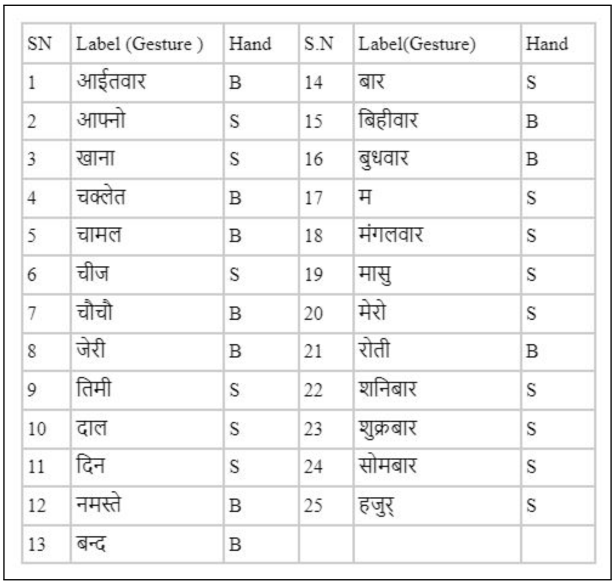


Figure 3: 4 Sample images of different individuals for gesture ‘Chocolate’

3.1.2 Description of signs

The gestures that are used are of different hand movements and there are variations in the hand gestures.The signs that have been used are listed below in the table. It includes the label name and another column specifies whether the sign.was performed using a single hand or both the hands. ‘S’ denotes that the sign was performed using a single hand and ‘B’ denotes that the sign was performed using both hands. The signs that have been used mostly belong to food items , week days and greetings which consists of variations in hand movements.

Table 1: Labels for different gestures



SN	Label (Gesture)	Hand	S.N	Label(Gesture)	Hand
1	आईतवार	B	14	बार	S
2	आफ्नो	S	15	बिहीवार	B
3	खाना	S	16	बुधवार	B
4	चक्लेत	B	17	म	S
5	चामल	B	18	मंगलवार	S
6	चीज	S	19	मासु	S
7	चौचौ	B	20	मेरो	S
8	जेरी	B	21	रोती	B
9	तिमी	S	22	शनिबार	S
10	दाल	S	23	शुक्रबार	S
11	दिन	S	24	सोमबार	S
12	नमस्ते	B	25	हजुर	S
13	बन्द	B			

3.1.3 Recordings

With proper consultation with Nepali Sign Language teacher, Mr. Hari Prasad Adhikari and with the help of the videos uploaded by him in his own youtube channel,different gestures were analyzed. All the recordings were done in different environmental conditions with natural lightning.It was taken in both outdoors as well as in indoor environments too. Subjects performed the signs standing with a plain background.All the 22 individuals use orange colored gloves in both the hands and recordings of the respective gestures have been performed.In order to remove all the issues associated with skin color variation, the individuals wore black colored dress and plain background.All the individuals stand in a plain background wearing orange colored gloves, then videos are recorded which are 1-2 secs long. The individuals are assisted by me in which the videos for NSL videos recorded by Hari Prasad Adhikari are also shown to the individuals.

The camera used for recording all the videos is Samsung Note 5. The resolution of the videos is 1920 by 1080 at 30 frames per second. Hence the videos are of good quality and the size of the data is a bit huge.Single gesture consists of 120 videos.Thus there are a total 3000 videos of 25 gestures.The total size of the dataset is 4GB and size of each video is 776kb but all the videos are not of same time, so the size of each video varies. Some videos are recorded for 1 second whereas some videos are recorded for 2 seconds as well depending upon the time taken to perform gesture.

3.2 Extraction of frames and preprocessing

A video is a collection of sequences of frames. Thus in order to process the videos ,respective frames contained in video should be extracted in order to get useful features of the particular gesture.We have used orange colored gloves that helps for proper color segmentation.The BGR values are noted down for the orange color consisting of the lower and upper boundary of it. We used colorpicker in order to find the BGR values of the gloves.Two masks are created then we perform bitwise ORing for these two masks in order to capture the required region of interest. The final preprocessed image that we get is a gray scale image of hands which acts as an input for the feature extractor module.

Two major steps are being carried out which consists of the feature extraction part and the classification part.The first approach in this nepali sign language gesture recognition consists of feature extraction by CNN model and classification part by RNN(LSTM) model in approach 1 and Vision Transformer in approach 2.

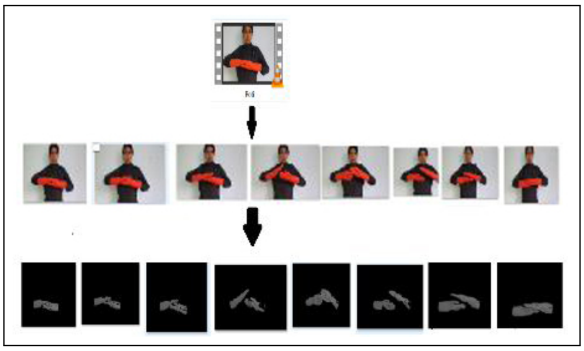


Figure 4: Extraction of frames from video and preprocessed images

3.3 Extracted frames given to CNN model

Three different architectures of CNN model are used for feature extraction in order to experiment what type of results do we obtain in these different architectures of CNN model.Inception V3, Densenet121 and VGG-16 are three different pretrained models which have been used here.

Two major steps are being carried out which consists of the feature extraction part and the classification part.The first approach in this nepali sign language gesture recognition consists of feature extraction by CNN model and classification part by RNN(LSTM) model in approach 1 and Vision Transformer in approach 2.

3.3.1 InceptionV3

The Inception V3 model is one of the pretrained models which has been trained with the ImageNet dataset consisting of 1000 classes. We will get 2048 features from the inception v3 model.

3.3.2 VGG-16

It is one of the pretrained CNN models which consists of 16 layers which have weights in it.There are 16 weighted layers in VGG16 where thirteen convolutional layers ,five max pooling layers , three dense layers and remaining 21 layers are present in it.

3.3.3 Densenet121

An architecture that mainly focuses on constructing the deep learning networks goes even deeper which consists of 120 convolutional layers and 4 average pooling layers.

3.4 Training model (temporal features)

Temporal features are those features that are highly associated with time or vary over time.Video consists of sequential frames in it which varies in time.Two approaches have been used for training the model for temporal features.

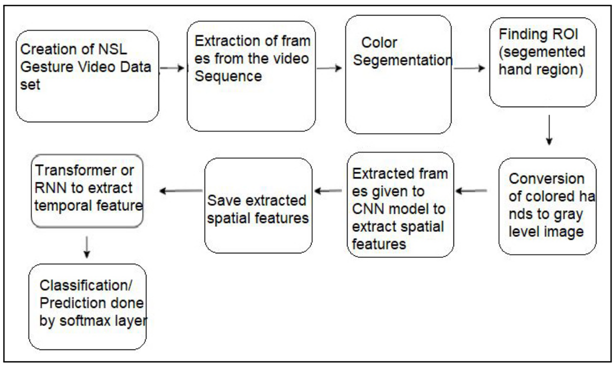

Creation of NSL Gesture Video Data set
Extraction of fram es from the video Sequence
Color Segementation
Finding ROI (segemented hand region)
Conversion of colored ha nds to gray level image
Extracted fram es given to CNN model to extract spatial features
Save extracted spatial features
Transformer or RNN to extract temporal feature
Classification/ Prediction done by softmax layer

Figure 5: System Block Diagram

3.4.1 Training RNN (Approach 1)

A Long Short-Term Memory (LSTM) model, which is an RNN with LSTM units is used for training purpose.The proposed model is a huge network which consists of single layer of 256 LSTM units.

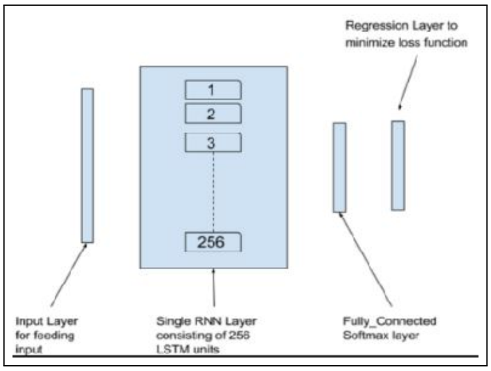
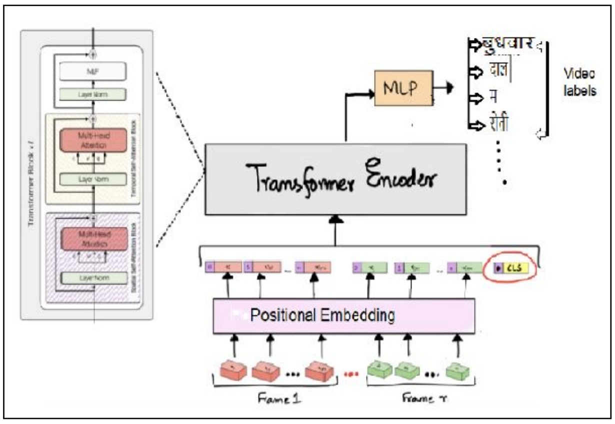

Regression Layer to minimize loss function
1
2
3
256
Input Layer for feeding input
Single RNN Layer consisting of 256 LSTM units
Fully_Connected Softmax layer

Figure 6: Used LSTM model

3.4.2 Training transformer (Approach 2)

A positional embedding layer is formed before transformer encoder where every tokens are passed through positional embedding layers and a position encoding is added to the encoded video tokens. Then all the tokens passes through the transformer encoder.The transformer encoder block is divided or broken down into two parts.The temporal self-attention layer computes attention between all tokens extracted from a different temporal index after the first multi-headed self-attention layer computes attention between all tokens extracted from the same spatial index (i.e., among all tokens extracted from the same clip).



An output label is obtained through a class token and the class tokens are fed to the feed forward neural network in order to get the final output prediction.

4. Evaluation Metrics

4.1 Cross entropy

The amount of ”wrong” or ”far away” the model’s forecast from the actual data determines cross entropy loss.

$$L_i = -\log\left(\frac{e^{f_{yi}}}{\sum_j e^{f_j}}\right) \tag{4.1}$$

$$H(p,q) = -\sum_x p(x)\log q(x) \tag{4.2}$$

$$\text{Where , } f_{yi}(z) = \frac{e^{zj}}{\sum_{k=1}^{C} e^{z_k}} \tag{4.3}$$

C=total number of classes. Equation 4.3 is the softmax function.

4.2 Prediction score

The softmax function is used for multi class classification which shows the mathematics behind the prediction score. The softmax equation can be given as

$$f_{yi}(z) = \frac{e^{zj}}{\sum_{k=1}^{C} e^{z_k}} \tag{4.4}$$

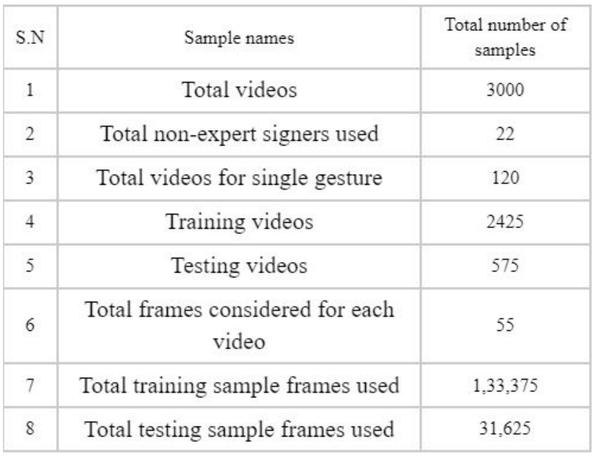
The Z here represents the data from the output layer’s neurons. Where C=total number of classes. The exponential is used to depict nonlinear functions. The sum of exponential values is used to standardize and transform these data into probabilities. The logit score matrix, often known as the net input matrix, is given by Z=XW+b where, X is a feature matrix of dimension n m, W is a weight matrix that represents the weight assigned to one feature for particular class label and has dimension m k b as the bias value.

5. Results and Discussion

There are altogether 3000 videos of nepali sign language gestures which consists of 25 classes that mostly consists of gestures for food items, weekdays, etc.80 percent of the total video sets are used for training and 20 percent are used for testing sets.Experiments have been performed on 25 types of Nepali sign language gestures. Each gesture consists of 120 videos each. Altogether 22 individuals have been used for acting about the gestures and they have performed five repetitions of each gesture.

The two models has been developed on python programming language, Python 3.8.3 on google colab.The model has been trained in NVIDIA RTX 2070.Various packages such as Opencv, Keras, Numpy, Matplotlib, Tensorflow,Transformers, etc have been used according to the need of the experiment. The table below shows the number of samples used in this experiment.

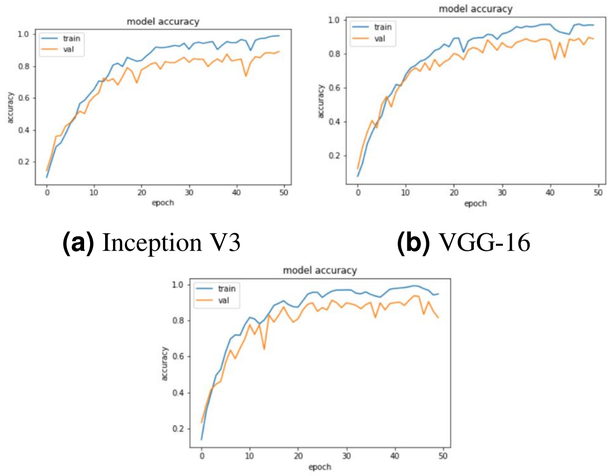
Table 2: Sample statistics



S.N	Sample names	Total number of samples
1	Total videos	3000
2	Total non-expert signers used	22
3	Total videos for single gesture	120
4	Training videos	2425
5	Testing videos	575
6	Total frames considered for each video	55
7	Total training sample frames used	1,33,375
8	Total testing sample frames used	31,625

5.1 Training RNN (Approach 1)

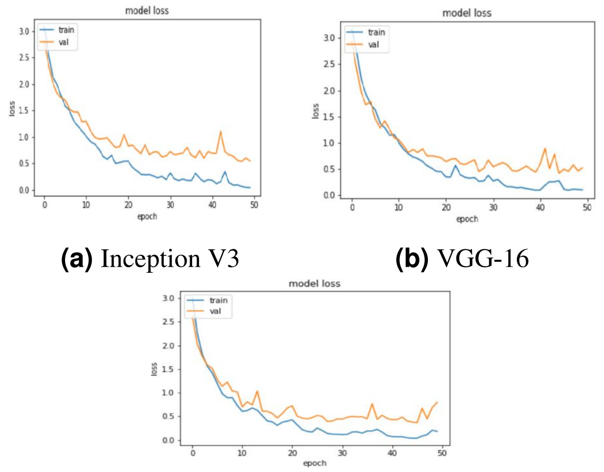
The experiment was carried out using three pretrained CNN architecture as a feature extractor and LSTM model has been used for training for temporal features. The convergence of model is found to be better in VGG-16



(a) Inception V3 **(b)** VGG-16

(c) DenseNet121

Figure 7: Train Vs Val. Accuracy for RNN



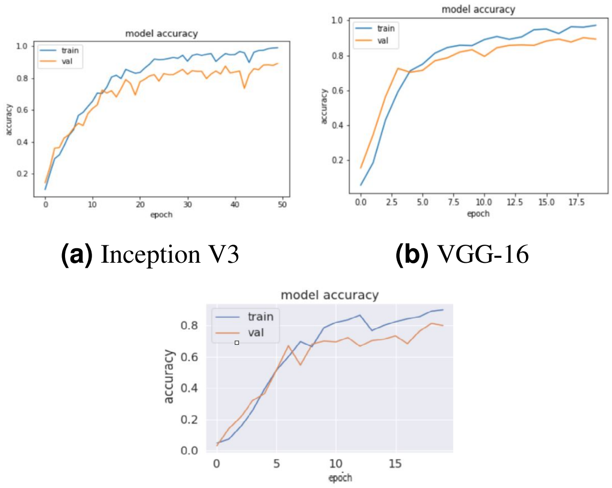
(a) Inception V3 **(b)** VGG-16

(c) DenseNet121

Figure 8: Train Vs Val. Loss for RNN

5.2 Training Transformer (Approach 2)

The experiment was carried out using three pretrained CNN architecture as a feature extractor and Vision Transformer model has been used for training for temporal features. The model convergence is seen better in VGG-16 using Vision Transformer .



(a) Inception V3 **(b)** VGG-16

(c) DenseNet121

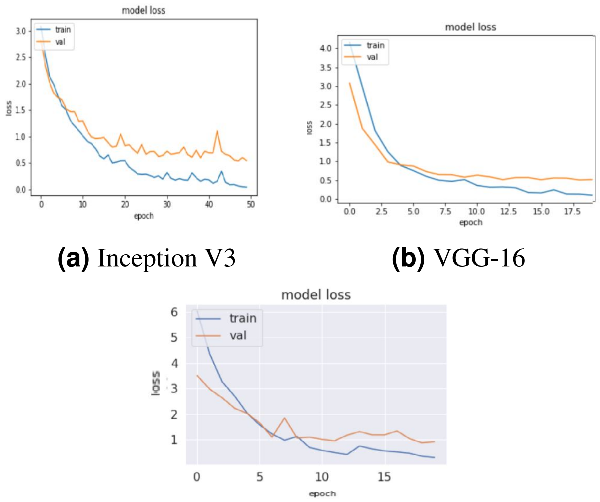
Figure 9: Train Vs Val. Accuracy for Transformer

5.3 Testing Trained Models

Total gestures used for testing =575 Total gestures per category=23

5.3.1 Testing RNN (Approach 1)

Out of 23 test videos following test videos are correctly predicted.



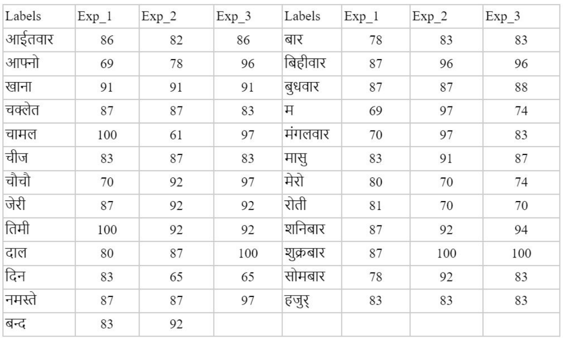
(a) Inception V3

(b) VGG-16

(c) DenseNet121

Figure 10: Train Vs Val. Loss for Transformer

Table 3: Accuracy per gesture using RNN(approach 1)

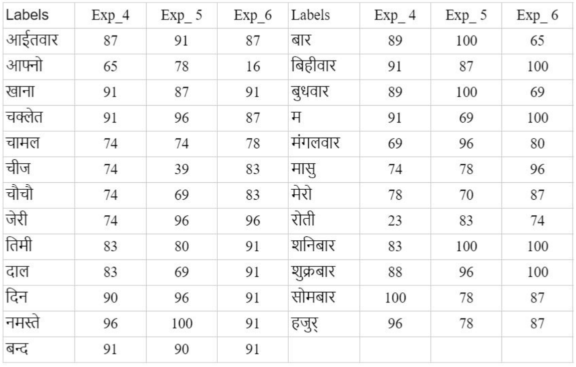


Labels	Exp_1	Exp_2	Exp_3	Labels	Exp_1	Exp_2	Exp_3
आईतवार	86	82	86	बार	78	83	83
आफ्नो	69	78	96	बिहीवार	87	96	96
खाना	91	91	91	बुधवार	87	87	88
चक्लेत	87	87	83	म	69	97	74
चामल	100	61	97	मंगलवार	70	97	83
चीज	83	87	83	मासु	83	91	87
चौचौ	70	92	97	मेरो	80	70	74
जेरी	87	92	92	रोती	81	70	70
तिमी	100	92	92	शनिबार	87	92	94
दाल	80	87	100	शुक्रबार	87	100	100
दिन	83	65	65	सोमबार	78	92	83
नमस्ते	87	87	97	हजुर्	83	83	83
बन्द	83	92					

5.3.2 Testing Transformer (Approach 2)

Out of 23 test videos following test videos are correctly predicted.

Table 4: Accuracy per gesture using Vision Transformer(approach 1)



Labels	Exp_4	Exp_ 5	Exp_6	Labels	Exp_ 4	Exp_ 5	Exp_ 6
आईतवार	87	91	87	बार	89	100	65
आफ्नो	65	78	16	बिहीवार	91	87	100
खाना	91	87	91	बुधवार	89	100	69
चक्लेत	91	96	87	म	91	69	100
चामल	74	74	78	मंगलवार	69	96	80
चीज	74	39	83	मासु	74	78	96
चौचौ	74	69	83	मेरो	78	70	87
जेरी	74	96	96	रोती	23	83	74
तिमी	83	80	91	शनिबार	83	100	100
दाल	83	69	91	शुक्रबार	88	96	100
दिन	90	96	91	सोमबार	100	78	87
नमस्ते	96	100	91	हजुर्	96	78	87
बन्द	91	90	91				

Table 5: Classification Results

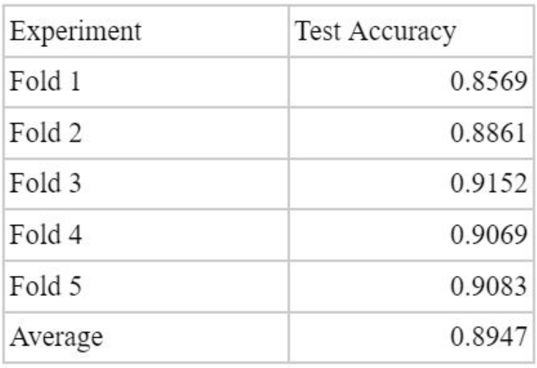
S.N	Pretrained CNN Model	Acc using RNN	Accuracy using Vision Transformer
1	Inceptionv3	79%	82%
2	VGG-16	87%	**88%**
3	DenseNet121	80%	86%

Among the three models used as a feature extractor, the best model was found to be a transformer using VGG-16 as a feature extractor. Thus, the second approach has an average test accuracy of 88

5.4 Validation of model

Kfold cross validation has been used in order to perform validation of the model. 5 fold cross validation has been done and the results can be shown in the table below:

Table 6: Kfold Test Results



Experiment	Test Accuracy
Fold 1	0.8569
Fold 2	0.8861
Fold 3	0.9152
Fold 4	0.9069
Fold 5	0.9083
Average	0.8947

Among the models used in six experiments, the best model was found to be using approach 2 in which VGG-16 has been used as a feature extractor module and Transformer has been used for classification . Thus, this model has been used while performing K-fold cross validation . The test accuracy of different folds can be shown in the table above in which the minimum test accuracy of the fold is found to be 0.8569 and maximum test accuracy of the fold is seen to be 0.9152. Thus the average accuracy of the model is 0.8947.

6. Conclusion

Thus, using two different trained models, the performance of transformer has found to be better with test accuracy of 88.Among the pretrained models used, VGG-16 is found to perform better with the classification models as compared to other two models that has been used for feature extraction. There are more features given as an output by VGG-16 as compared to other models because of which the result is better as more features have been extracted. Thus, performance in approach 2 is found to be better as compared to the performance in approach 1

Hence, the transformer is found to be performing better as compared to LSTM. All the inputs have been ingested at once by using a transformer whereas all the input feature vectors are not ingested at once using LSTM because of which there is no parallel processing in approach 1.This accounts for better performance by using approach 2.

The task of hyperparameter tuning for different models in order to increase the accuracy of the models can be done.

Acknowledgments

The authors acknowledge all the friends and family members who assisted for dataset collection and video recordings. Authors are also thankful to the Department of Electronics and Computer Engineering of IOE Pulchowk Campus for valuable guidance and advice.

References

[1] Vivek Thapa, Jhuma Sunuwar, and Ratika Pradhan. Finger spelling recognition for nepali sign language. In *Recent Developments in Machine Learning and Data Analytics*, pages 219–227. Springer, 2019.

[2] Drish Mali, Rubash Mali, Sushila Sipai, and Sanjeeb Prasad Panday. Two dimensional (2d) convolutional neural network for nepali sign language recognition. In *2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*, pages 1–5. IEEE, 2018.

[3] Anup Nandy, Soumik Mondal, Jay Shankar Prasad, Pavan Chakraborty, and GC Nandi. Recognizing & interpreting indian sign language gesture for human robot interaction. In *2010 international conference on computer and communication technology (ICCCT)*, pages 712–717. IEEE, 2010.

[4] Rung-Huei Liang and Ming Ouhyoung. A real-time continuous gesture recognition system for sign language. In *Proceedings third IEEE international conference on automatic face and gesture recognition*, pages 558–567. IEEE, 1998.

[5] Ankita Saxena, Deepak Kumar Jain, and Ananya Singhal. Sign language recognition using principal component analysis. In *2014 Fourth International Conference on Communication Systems and Network Technologies*, pages 810–813. IEEE, 2014.

[6] Jung-Bae Kim, Kwang-Hyun Park, Won-Chul Bang, and Z Zenn Bien. Continuous gesture recognition system for korean sign language based on fuzzy logic and hidden markov model. In *2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE'02. Proceedings (Cat. No. 02CH37291)*, volume 2, pages 1574–1579. IEEE, 2002.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Sarfaraz Masood, Harish Chandra Thuwal, and Adhyan Srivastava. American sign language character recognition using convolution neural network. In *Smart Computing and Informatics*, pages 403–412. Springer, 2018.

[9] Sarfaraz Masood, Adhyan Srivastava, Harish Chandra Thuwal, and Musheer Ahmad. Real-time sign language gesture (word) recognition from video sequences using cnn and rnn. In *Intelligent Engineering Informatics*, pages 623–632. Springer, 2018.

The dataset link for NSL Gesture video:
https://drive.google.com/file/d/1ygPcev5c8JssuwOr4KEqPfBNqn81OGat/view?usp=sharing