Extraction of Motion Trajectories and Evaluation of Gesture Recognition Accuracy from Video Data

Anderson Nguyen[1] , Vangelis Metsis[2]

Department of Computer Science, Texas State University, San Marcos Texas

nguyenande@gmail.com

Abstract

This research implements a performance metric system aimed to evaluate gesture recognition accuracy. We utilized a variety of novel deep learning models to assess how accurate a machine can classify a label from videos that contains motion trajectories in forms of gesture and sign language. In addition, we implement automatic scripting for landmarks coordination extraction using holistic tracking of face, body, and hands.

Goals: During this REU program, the goal is to determine the classification accuracy of different deep learning architectures using extracted data from sign language videos.

Introduction

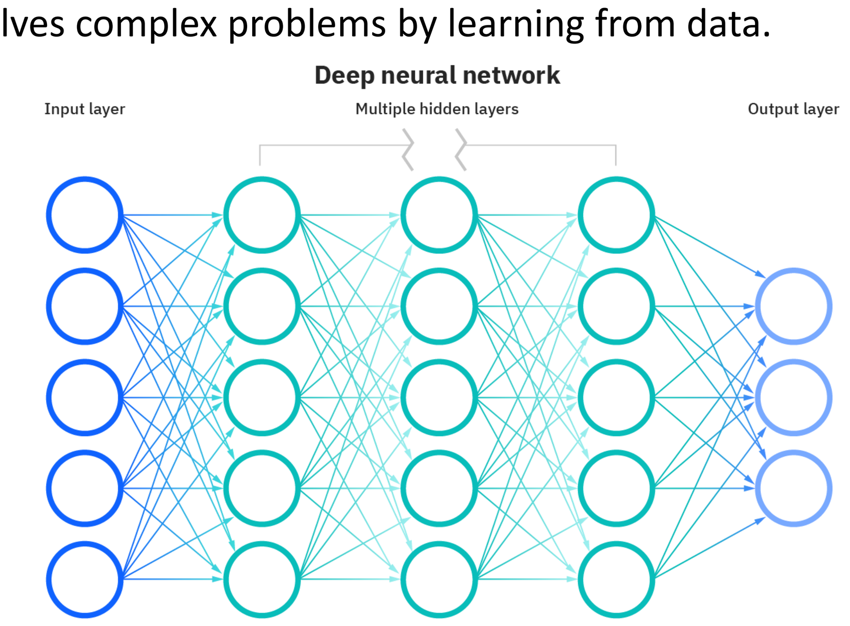
Motivation: Videos provide a vast amount of information. What if we can harness every piece of data in the video and feed it to a machine for evaluation. What could we do with this data? More importantly, how do we determine the accuracy of this data and what future applications can this benefit?

Potential Applications:

- Robotic Vision
- Hand Gesture Control
- Sign Language Recognition
- Motion Capture
- Virtual and Augmented Reality

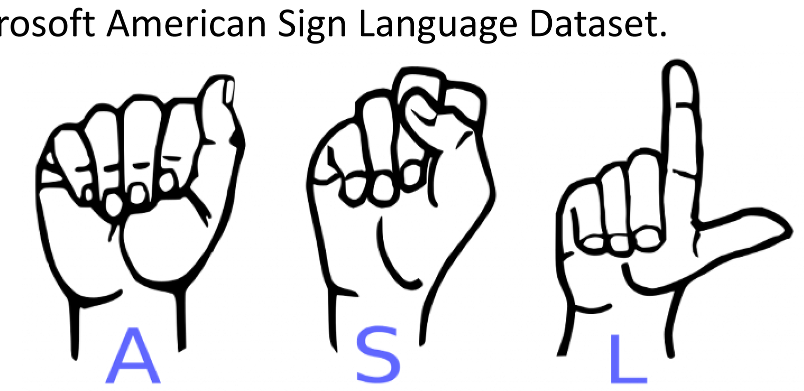
Deep Learning:

- Machine learning based on artificial neural networks.
- Multiple layers of processing.
- Collect, analyze, and interpret large amounts of data.
- Solves complex problems by learning from data.



Our Dataset:

- Data extracted from personally created videos.
- American Sign Language Lexicon Video Dataset.
- Microsoft American Sign Language Dataset.



Methodology

Frameworks and Tools:

- Python
- Google MediaPipe
- OpenCV
- Numpy
- Panda
- SkLearn
- Keras/TensorFlow
- Matplotlib

MediaPipe OpenCV Pandas NumPy

Deep Learning Architectures:

- Long Short-Term Memory Neural Network (LSTM)
- LSTM with Self-Attention
- Transformers
- 1D-Convolution Neural Network (CNN)

Process:

Step 1: Create or choose a dataset and prepare for training.
Step 2: Assign labels and features.
Step 3: Split the dataset in test and train subsets.
Step 4: Normalize X and labels to categorical data.
Step 5: Split X and Y.
Step 6: Compile and train the deep learning model.
Step 7: Obtain accuracy, loss, and classification score of model.
Step 8: Repeat above steps with all four deep learning models.

Figure: Landmark detection on face, body, and hands using Google MediaPipe Holistic for data extraction.

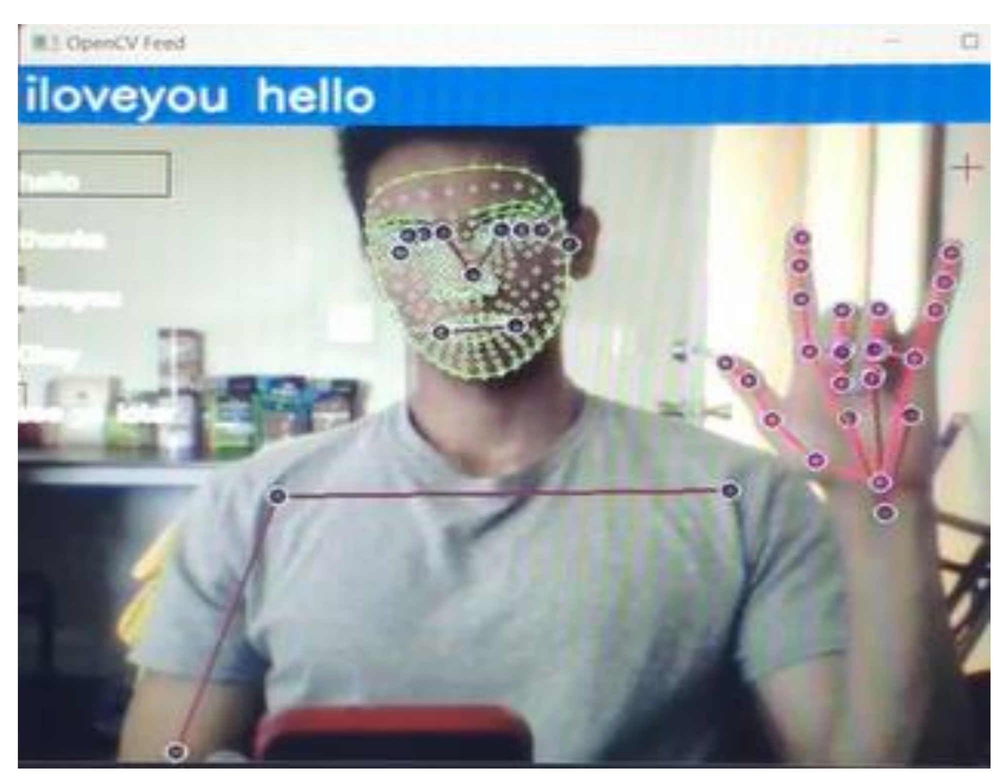
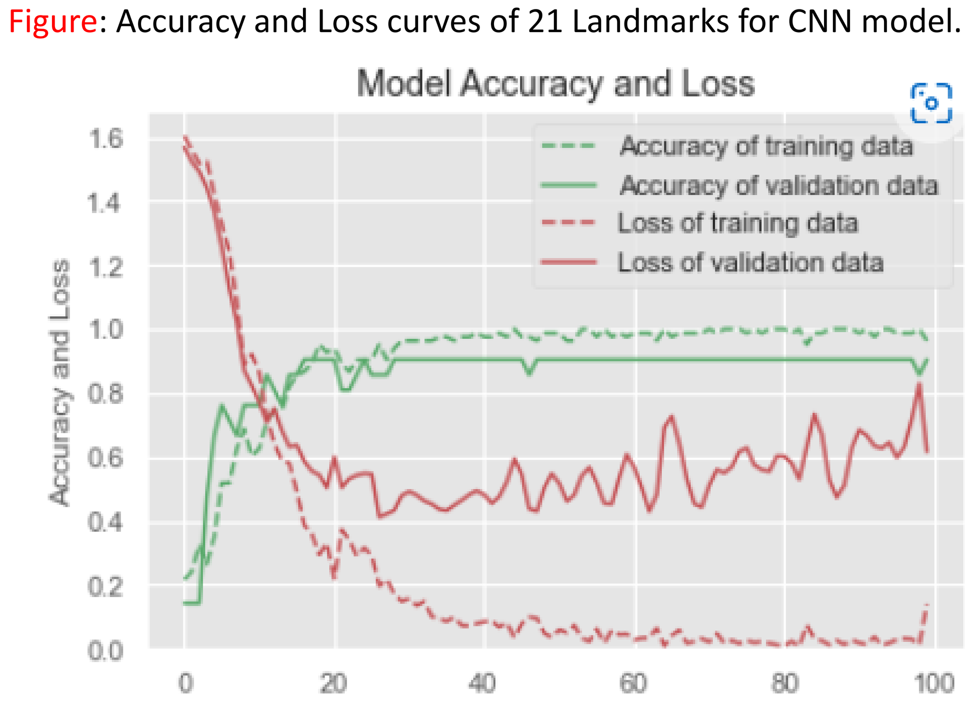


Figure: Accuracy and Loss curves of 21 Landmarks for CNN model.



Result

Figure: Classification report for CNN test data via confusion matrix

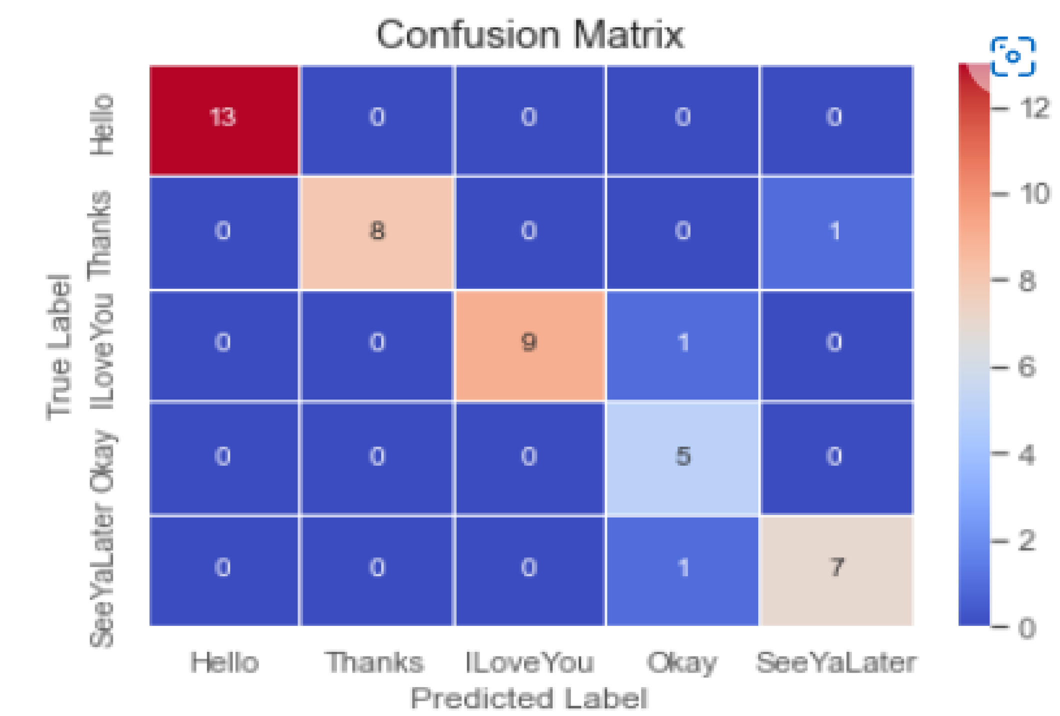


Figure: Accuracy score bar graph. Accuracy score is the number of correct classification obtained over total classifications.

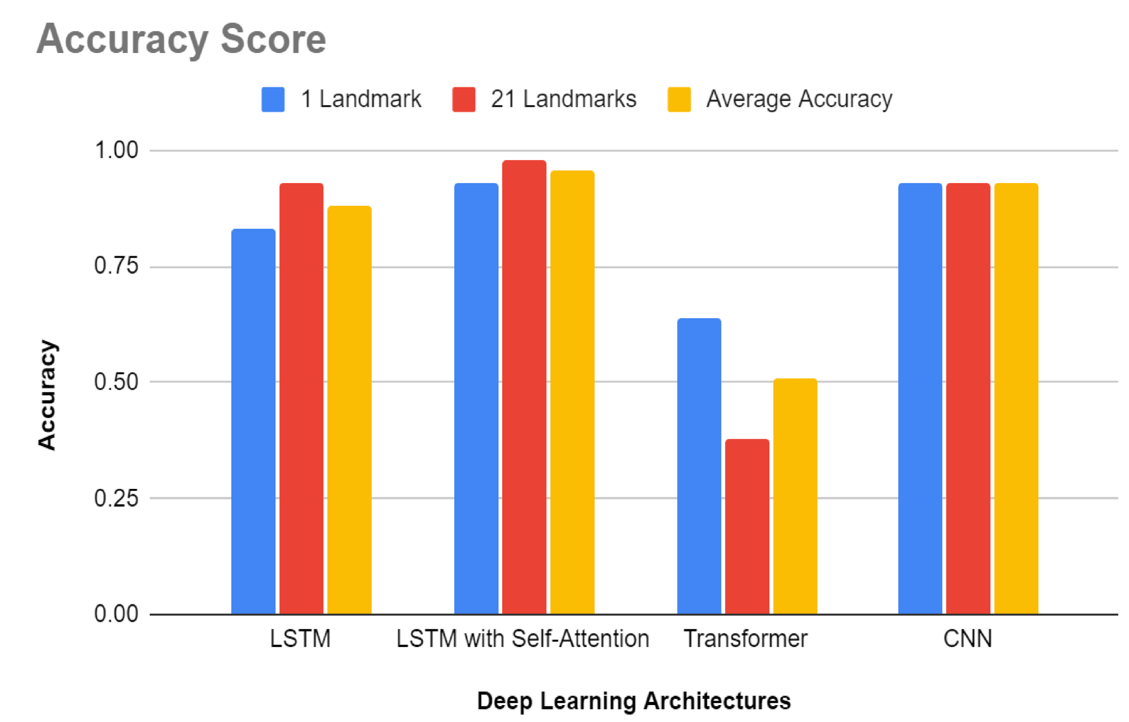
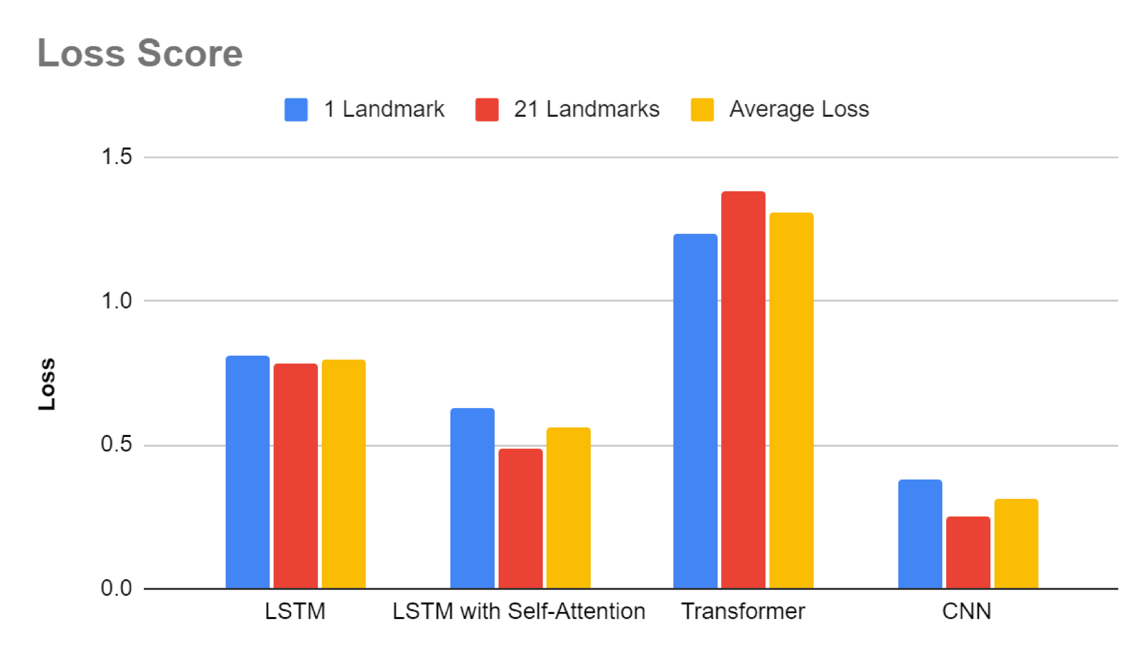


Figure: Loss score bar graph. Loss is a value that represents the summation of errors in the model.



The dataset utilized in this result were personally created. They consisted of 5 sign labels ("hello", "thanks", "I love you", "okay", "see you later"). A correct classification refers to a model predicting a sign label correctly when given a sign without a label. This is known as the accuracy score. Lost value implies how well or poorly a certain model behaves after each iteration of optimization. This is a value we want to minimize and is also known as Loss Score.

Conclusion

Full Comparison Analysis:

LSTM (Self-Attention)

- Less data required hence faster to train.
- Highest average accuracy.
- Performs better at higher number of landmarks.

LSTM with self-attention layer perform slightly better than stand alone LSTM in terms of accuracy and loss score.

Transformers:

- Lowest average accuracy.
- Performs better with one landmark.
- Requires more data to train model successfully.

One interesting finding is transformer model outperformed all other models when extracted data has been randomized before being feed into neural network.

1D-Convolution Neural Network:

- Lowest loss.
- Consistent high accuracies between one and 21 landmarks.

Areas for future research:

- 543 total landmarks from face, body, and hands instead of 1 and 21 landmark(s) from right hand.
- Utilize larger dataset to test and train existing deep learning architecture.
- Cross Validation.
- Capture and classify complete sentences.
- Implement semi supervised machine learning.
- Real time detections, classifications, and predictions.

Acknowledgements

This material is based upon work supported by the National Science Foundation under REU grant #1757893. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan and A. Thangali, The ASL Lexicon Video Dataset, CVPR 2008 Workshop on Human Communicative Behaviour Analysis (CVPR4HB'08) (pdf ps)

https://google.github.io/mediapipe/

https://github.com/nicknochnack/ActionDetectionforSignLanguage

TEXAS STATE UNIVERSITY

The rising STAR of Texas

NSF