SignLens: A Sign Language Translation Model for Specially Abled

Gagan Singhal, Ananya Singh, Andrew Wood
gsinghal@bu.edu, xananya@bu.edu, aewood@bu.edu
Boston University, MA, USA

Introduction

Communication barriers between the deaf and mute and those unfamiliar with sign language remain a significant challenge in today's world. We propose a real-time translation system that converts American Sign Language into text to address this. Utilizing the WLASL dataset, the system captures the intricate nuances of ASL through a multimodal architecture: one model processes video frames to recognize individual signs. At the same time, another refines these into meaningful text by leveraging contextual word relationships. By combining Computer Vision and Natural Language Processing, our solution aims to facilitate seamless communication across educational, professional, and social settings, enabling broader accessibility.

DATASETS	GLOSS	VIDEOS	SIGNERS
WLASL100	100	2038	97
WLASL300	300	5117	109
WLASL1000	1000	13168	116
WLASL2000	2000	21083	119

Figure 1: Human Pose Estimation using OpenPose

Motivation

Hearing loss affects 466 million people worldwide, or 6.1% of the population, making it the fourth leading cause of disability. Projections suggest this number could exceed 900 million by 2050. Notably, 95% of children with hearing disabilities are the first in their families to experience such conditions, leading to significant challenges in communication and learning. This not only affects the individuals but also profoundly impacts their families and communities, often resulting in social isolation and limited opportunities for growth.

- Bridge the gap between sign language users and non-users.
- Enable instant communication in live settings (e.g., conferences, classrooms) by generating meaningful, grammatically correct sentences.
- Enhance accessibility places like airports, hospitals, etc.
- Enable interaction with AI and assistive technologies.
- Empower the deaf with greater autonomy in daily life.

Objectives

Use the WLASL dataset, which contains 2,000 vocabulary items to do the following:

- Create a model that effectively gives good results without using pre-built models like I3D and TGCN and still be comparable to them.
- Evaluate the model's performance using the top-K classification accuracy mean scores with $K = 1, 5, 10$.
- Assess the models over all the sign instances on a subdivided dataset, where the dataset is divided into top N glosses, with $N = 100, 300, 1000, 2000$.
- Stack the above model over an NLP model to stitch the predicted words and effectively generate meaningful sentences.

Baseline

The baseline model evaluates the performance of Pose-TGCN and I3D, with Pose-TGCN focusing on motion dynamics from pose data and I3D capturing spatiotemporal features from raw video frames.

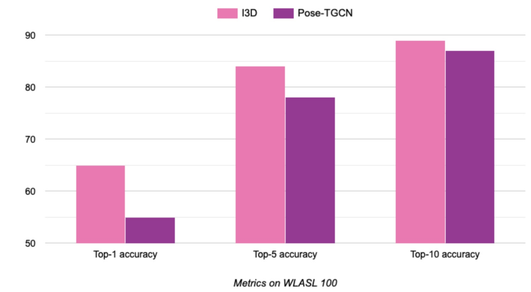

Figure 2: Top K classification accuracy percentage of baseline models

- Pose-TGCN: Despite being smaller in size than the I3D model, it achieves comparable tap-5 and top-10 accuracies on the WLASL100 and even the WLASL2000 subset, demonstrating its effectiveness in encoding human motion. However, its reliance on external pose estimator can lead to performance degradation due to pose estimation errors.
- In contrast, I3D processes row video data end-to-end, reducing spatial feature errors during training. This suggests that incorporating end-to-end training for pose-based models like Pose-TGCN could further improve their recognition performance

Model Details

- **Spatial Feature Extraction:** MobileNetV2 (CNN)
- **Temporal Sequence Learning:** GRU-based RNN
- **Classification:** Fully connected layers and softmax function to transform temporal features into a probability distribution over potential ASL signs, producing the final prediction.
- **Sentence Formation:** N-Gram to predict the next most probable word.

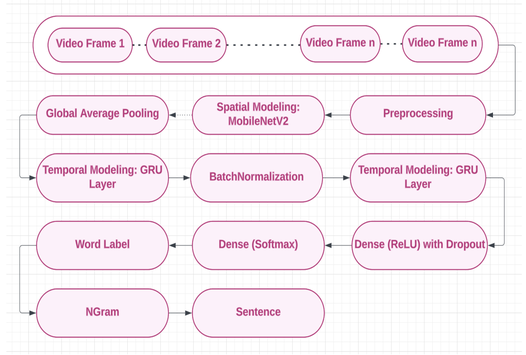

Figure 3: Workflow of proposed model for sign language recognition

- **Feature Extraction: MobileNetV2 (CNN)**

$$\mathbf{F}_t = \text{Conv}(\mathbf{I}_t, \mathbf{W})$$

where $\mathbf{I}_t$ = frame at time t, and $\mathbf{W}$ = learned kernel.

- **Temporal Sequence Learning: GRU (Recurrent Neural Network)**

$$\mathbf{h}_t = \mathbf{z}_t\mathbf{h}_{t-1} + (1-\mathbf{z}_t)\tanh(\mathbf{W}_h\mathbf{x}_t + \mathbf{U}_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h)$$

where $\mathbf{r}_t$=reset gate, $\mathbf{z}_t$ =update gate, and $\mathbf{h}_t$=output hidden state encoding frames' sequential relationships.

- **Classification: Fully Connected Layers and Softmax**

$$\mathbf{y} = \text{Dense}(\mathbf{h}_T) \quad P(y_i|\mathbf{x}) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

$$\hat{y} = \arg\max(P(y_i|\mathbf{x})) \quad \text{where } z_i = \text{logits}$$

- **Sentence Formation: N-Gram**

$$P(w_n|w_1, w_2, ..., w_{n-1}) = \frac{\text{Count}(w_1, w_2, ..., w_{n-1}, w_n)}{\text{Count}(w_1, w_2, ..., w_{n-1})}$$

$$\hat{w}_n = \arg\max_{w_n} P(w_n|w_1, w_2, ..., w_{n-1})$$

Observations

After training the model on the WLASL100 dataset for 300 epochs and 12 hours of training, we calculated the top k scores for k = [1, 5 10] and calculated the precision, recall, and F1 scores. We achieved an accuracy of about 23% on the test set.

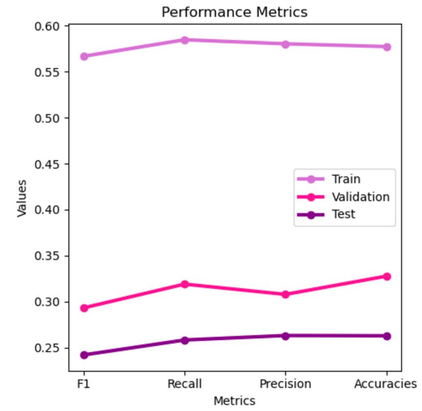

Figure 4: Performance results for training, validation and testing sets

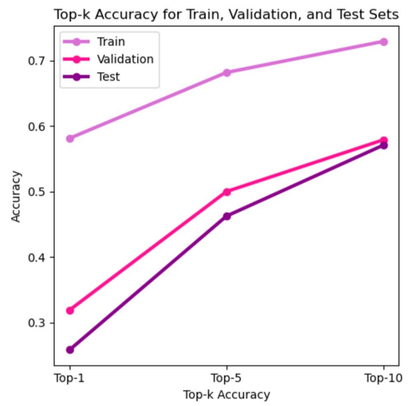

Figure 5: Top K accuracy for training, validation and testing sets

Conclusion

We found that even using a traditional approach to sign language translation could produce acceptable results if you have enough resources. Using state-of-the-art models with enough resources to fine-tune the model or developing more advanced and diverse datasets can actually help us make an efficient model that could tackle this problem.