Sign Pose-Based TRANSFORMER For Word-Level Sign Language Recognition [SPOTER]

Rohan Anil Gupta [M. Eng. - Computer Science - Blacksburg Campus]
Vinit Anishkumar Masrani [M. Eng. - Computer Science - Blacksburg Campus]
Saurav Kumar [MS - Computer Science - Blacksburg Campus]

INTRODUCTION

- Author’s focus is on **Manual** and **Isolated** Sign Language Recognition (SLR)
- Evaluation of the Model is Done using **Transformers**
- Transformer models are **computationally cheap** and have **outstanding performance** in **sequential tasks**
- Main Contributions are:
  - Constituting state of the art results on the WLASL-100, WLASL-300, and LSA64 datasets when considering pose-based SLR.
  - Novel normalization scheme.
  - Sequential joint rotation augmentation of the body pose.
  - Analysis of the pose-based vs appearance-based approaches.

METHODOLOGY - SPOTER (Sign POse-based TransformER)

A) Pre - Processing

- Obtain Pose - Estimates
- Extract 54 Landmarks
- 5 Head Landmarks
  - Head Landmarks include 2 eyes, 2 ears and the nose
- 21 Landmarks per hand
  - Hand Landmarks include (4 joints per finger) and one joint for wrist
- All these 54 Landmark are 2D
- We thus obtain 108 dimensional pose vector per frame

METHODOLOGY - SPOTER (Sign POse-based TransformER)

B) Augmentations

- In Plane Rotations

$$f_{\text{rotate}}(x, y) = ((x - 0.5)\cos\theta - (y - 0.5)\sin\theta + 0.5,\\ (y - 0.5)\cos\theta + (x - 0.5)\sin\theta + 0.5),$$

- Squeeze

$$f_{\text{squeeze}}(x) = \frac{x - w_1}{W - (w_1 + w_2)},$$

- Perspective Transformation
- Sequential joint Rotation

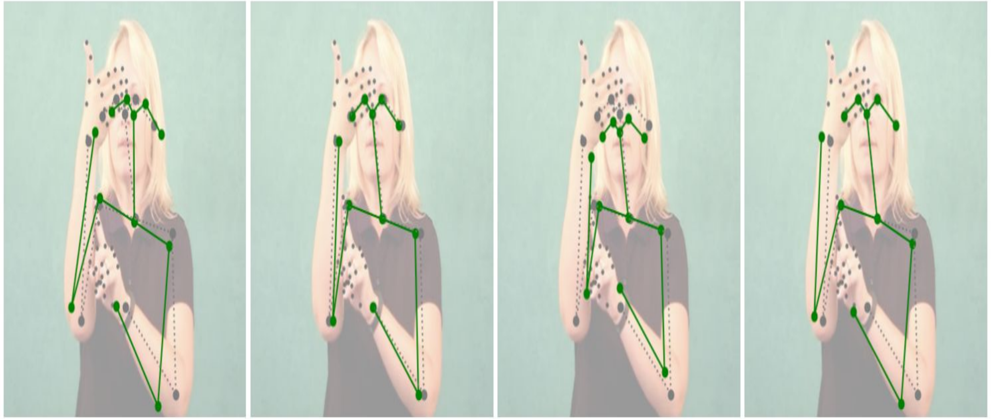
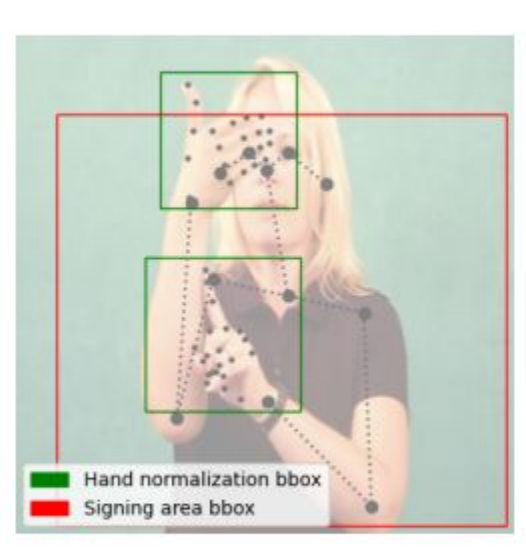


Figure 1. Depiction of individual augmentations applied on single frames. From left to right, there is in-plane rotation, squeeze, perspective transformation, and sequential joint rotation augmentation.

METHODOLOGY - SPOTER (Sign POse-based TransformER)

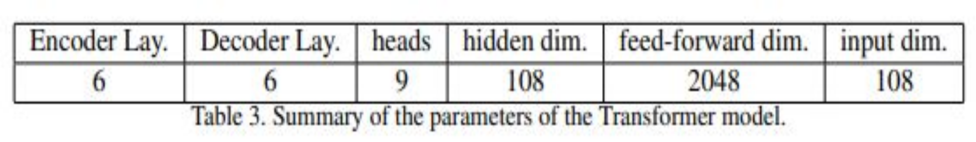
C) Normalization

- Purpose
- Signing Area Box
- Hand Pose Landmark Normalization Box
- Shift the Final Box



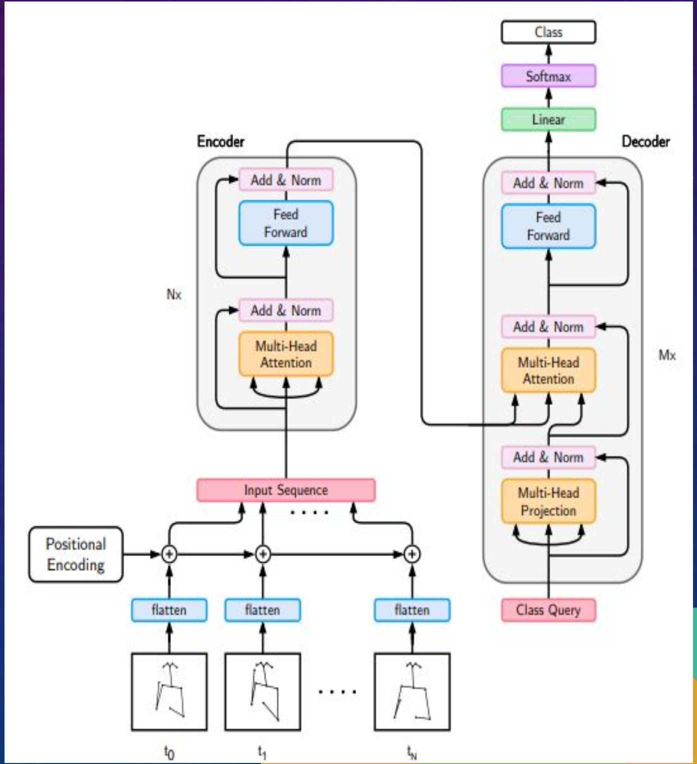
METHODOLOGY - SPOTER (Sign POse-based TransformER)

D) Proposed Architecture



Encoder Lay.	Decoder Lay.	heads	hidden dim.	feed-forward dim.	input dim.
6	6	9	108	2048	108

Table 3. Summary of the parameters of the Transformer model.



EXPERIMENTAL ANALYSIS

Implementation Details:

- The proposed architecture has been implemented in:
  - *PyTorch*

- Summary of parameters used for reproducing the proposed results:

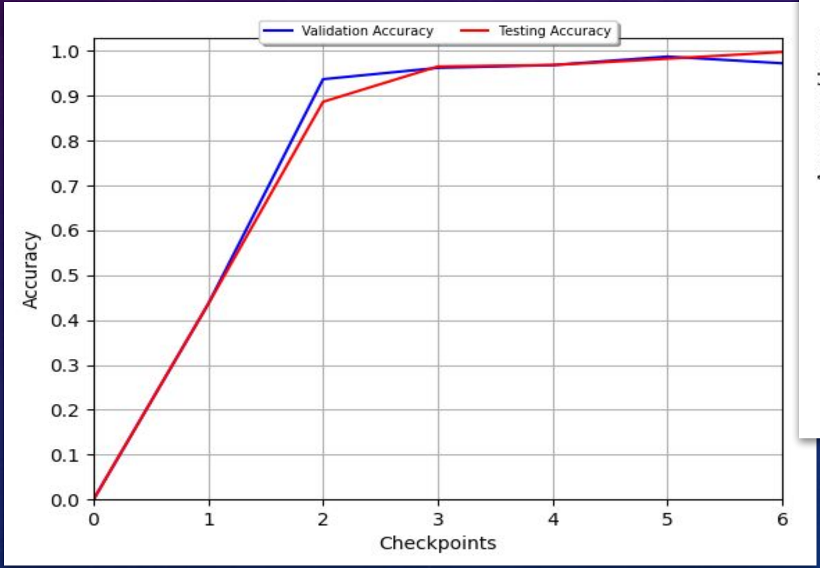
	Epoch	Learning Rate	Loss Function	Initial Weight Distribution	Validation Set Size
Proposed	350	10^{-3}	Cross Entropy Loss	Uniform over [0, 1]	20%
Reproduced	100	10^{-3}	Cross Entropy Loss	Uniform over [0, 1]	20%

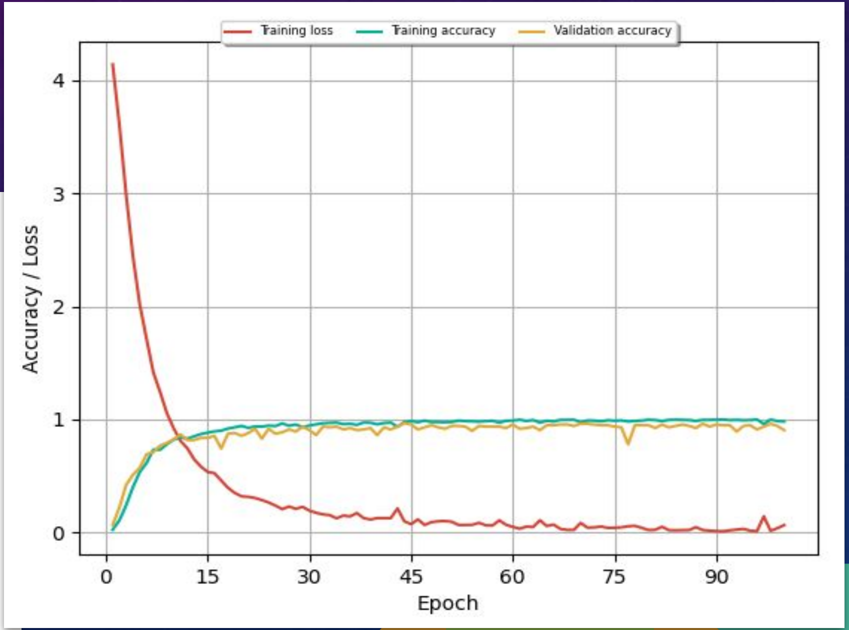
EXPERIMENTAL ANALYSIS - Performance on LSA64

- LSA64 Dataset:
  - 64 Classes, 3200 Instances
- Model is fed pre-processed data (open-source):
  - 64 Classes, 3178 Instances
  - Dataset generation:
    - Single Random split of pre-processed data:
      - Training & Validation - 80 % [2542 data-points]
      - Testing - 20 % [636 data-points]
    - Single Random split of training & validation data:
      - Training - 80 % [2034 data-points]
      - Validation - 20 % [508 data-points]

EXPERIMENTAL ANALYSIS - Performance on LSA64

- Baseline - IED Model (98.91 %)
- Proposed State of the art Test Accuracy - 100 %
- Reproduced Test Accuracy - 99.75 %



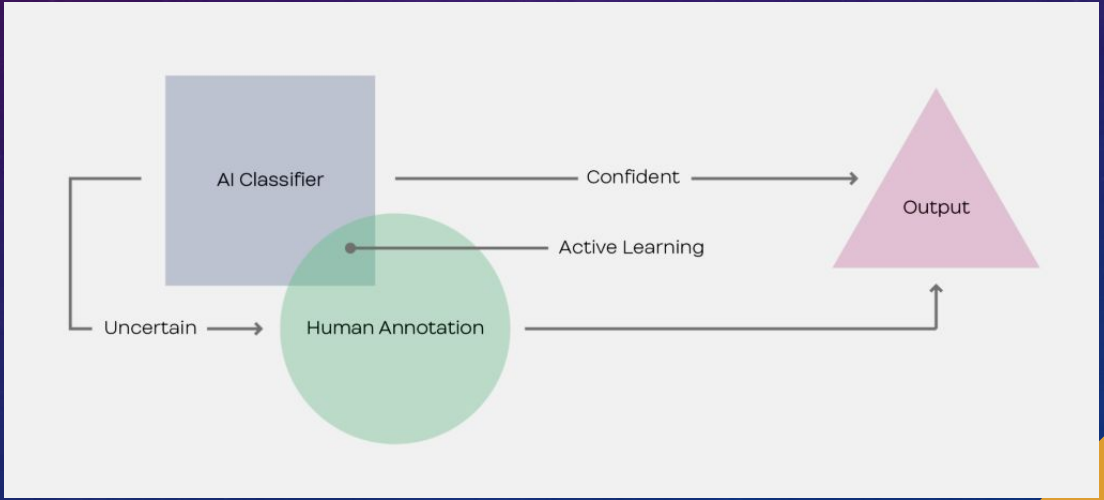


EXPERIMENTAL ANALYSIS - Performance on WLASL

- WLASL Dataset:
  - 2000 Classes, 21083 Instances
- Model is fed pre-processed data (open-source):
  - 100 Classes, 2037 Instances
  - Datasets:
    - Training: 70 % [1442 data-points]
    - Validation: ~17 % [337 data-points]
    - Testing: ~13 % [258 data-points]
- Unfortunately, our team wasn't able to reproduce learning on WLASL dataset. We could not run the source code due to runtime IndexErrors.
  - We are corresponding with the author (Link)

FUTURE WORK

- Use of SPOTER in **Human In The Loop(HITL)** fashion
- Training an **Appearance** based model



SUMMARY

- Use of Transformer for SLR
- Significant Improvement over previous works
- Creation of new data augmentation technique specific for SL
- Validated approach on 2 datasets
- Results reproduced only on 1 dataset

Lessons we learned

- Use and power of transformers in AI/ML
- Pose Estimation from Sign Language Actions
- The application of augmentation and normalization in sign language recognition
- Investigate the reproducibility of the paper
- Gained knowledge about different types of sign languages

ACKNOWLEDGEMENT

- This project is based on the work of *Matyas Bohacek* and *Marek Hruz*, University of West Bohemia, Czech Republic [PDF]
  - DOI: **10.1109/WACVW54805.2022.00024**

- Thanks to Dr. Yinlin Chen for providing the opportunity to explore such interesting concepts!

THANK YOU!

QUESTIONS?