Tracking ASL Gestures Using Autoencoders

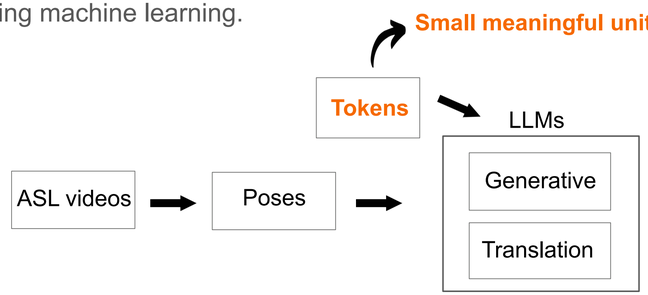
Katyani Mehra (km7872@rit.edu) | Advisor: Richard Lange

RIT | Golisano College of Computing and Information Sciences

Computer Science @ RIT

Introduction

Sign languages use **visual** elements like **handshape**, **location**, and **movement** instead of spoken words. These elements, or **tokens**, are essential for recognition, translation, and generation using machine learning.



Motivation

Current ASL translation systems often treat phrases as **English to Sign** mappings, instead of **True ASL**.

Current systems fail to capture ASL grammar

A major reason for this difference is the **lack of effective tokenization**. **Manual tokenization** for ASL videos is **time-consuming** and **costly**. To address this, we propose using **Vector Quantised - Variational AutoEncoder (VQ-VAE)**, an **unsupervised** model that **compresses** ASL videos into **discrete tokens**.

Why VQ-VAE?

VQ-VAE's excel at working with **discrete data**, making them well-suited for sign language, where tokens don't map directly to individual frames or body parts:

- **1 token ≠ 1 frame**
- **1 gesture ≠ 1 body part**

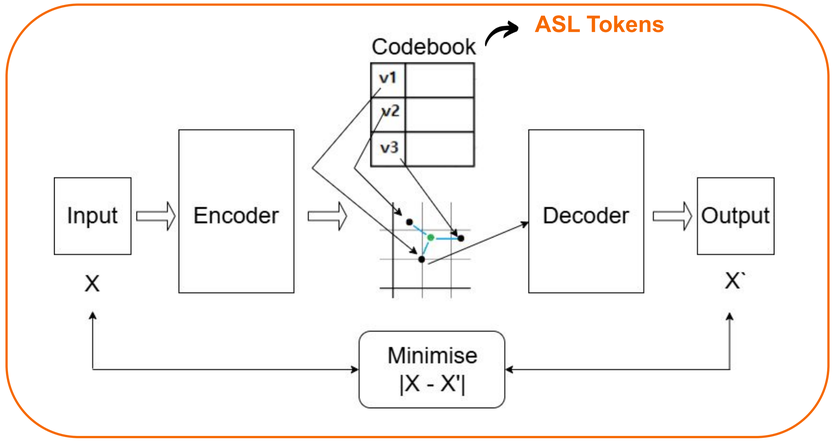
Approach

MS-ASL Dataset

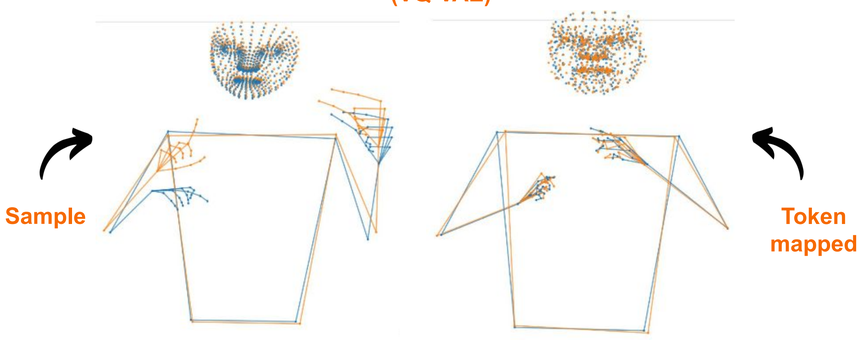
- **Continuous** ASL phrases.
- 456 videos = ~ 1900 samples = **~ 900000 frames**

Mediapipe Holistic Pose Estimation Model

- **Real-time** human pose detection model.
- Tracks: Body, right hand, left hand, face landmarks
- Each frame has **543** Keypoints (**3D**).



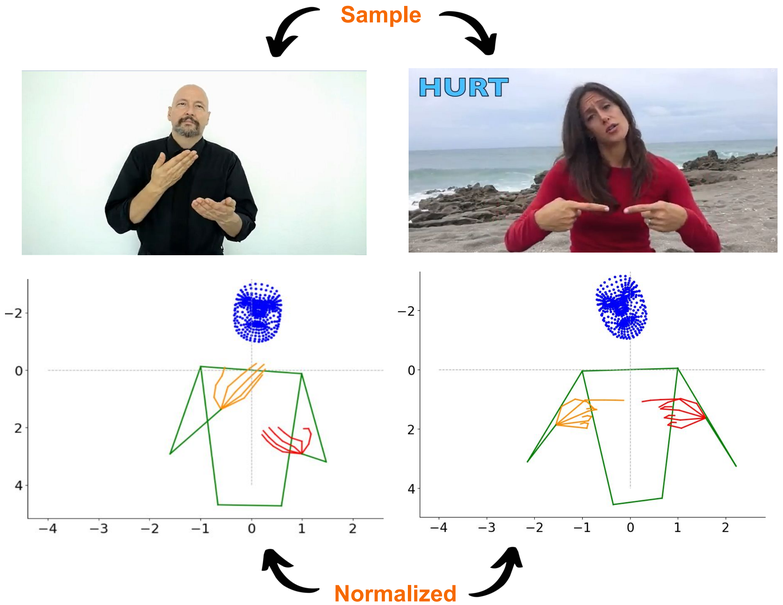
Vector Quantised - Variational AutoEncoder (VQ-VAE)



Sample to Token mapping

Normalization

To account for variations in size and position due to camera distance and frame placement, we normalize landmarks to a **new coordinate system**. In this system, the **left shoulder** is mapped to **1**, the **right shoulder** to **-1**, and the **center** of the shoulders is set to **0**. The body is then **scaled** to a consistent size by calculating the distance between the left and right shoulders and adjusting the scale accordingly.



Results

The mutual information score is 1.03, indicating a strong but not perfect dependency between the ASL gestures and the tokens produced. With further training, we expect to improve this score and generate more tokens.

References

- Google AI, "Holistic landmarks detection task guide", 2024. Available at: https://ai.google.dev/edge/mediapipe/solutions/vision/holistic_landmarker.
- "Neural Discrete Representation Learning", [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.