Gaze and Head Joint Representation Learning

A Multimodal Approach

Ashwin Murali, Souptik Kumar Majumdar

February 7, 2024

Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

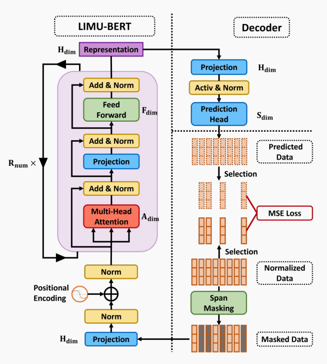
Supervisor: Chuhan Jiao

Table of Contents

- Accurate capture and interpretation of gaze data in digital interaction faces hurdles.
- Frequent missing data poses challenges, stemming from calibration issues, human errors, and natural actions like blinking.
- These challenges can be tackled by using head movement data to enhance and rebuild incomplete gaze data, improving accuracy in digital interactions and analysis.

Background

- LIMU-BERT aims to use unlabeled IMU data to extract generalized features, adopting self-supervised training principles in IMU sensor measurements.



Source: LIMU BERT Architecture. (Xu et al. [2021])

Objectives

- Extend LIMU-BERT model as a Multi-Modal, aiming to leverage correlations between head and gaze for improved gaze data reconstruction.
- Examine the multimodal performance on challenging test sets and downstream tasks to assess how this correlation contributes to the reconstruction process.

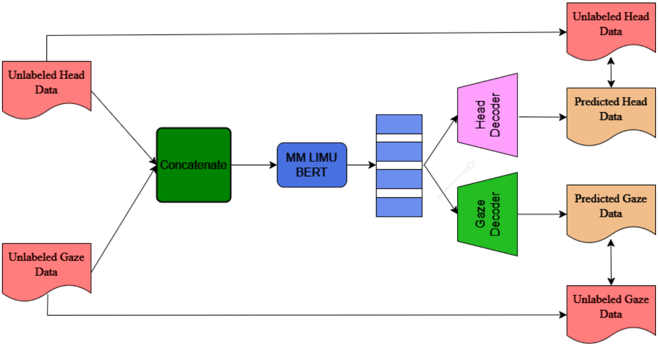
Feature	Description
Name	VR Behavior (V2) (Jin et al. [2022])
Modalities	Head and Gaze
Data Representation	Coordinates and quaternions
Sampling Frequency	120Hz among 100 users
Total Samples	60,926

- Unit vectors of head and gaze representing the direction are the input features for the Model.
- Input sequences of 8 seconds each are provided to the model.
- **Data Preprocessing**
  - The data was downsampled from 120Hz to 30Hz.
  - Rows containing NaN values were removed.

Masking

- The Masked Language Model (MLM) task involves masking individual tokens in a sequence and training the model to predict the masked tokens based on the surrounding context.
- Because adjacent IMU sensor data measurements are similar over time, the model can reconstruct the masked readings by mirroring neighboring readings.
- To provide a sufficiently challenging condition for training an effective model, longer subsequences are masked instead of just one token using span masking algorithm.

Multi Modal Architecture



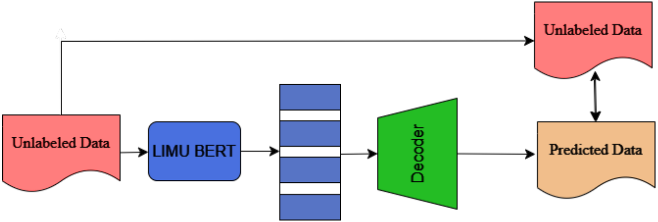
Source: Multi Modal LIMU BERT Architecture

Variants

- Multi Modal with Gaze Reconstruction
- Multi Modal with Head and Gaze Reconstruction

Baselines

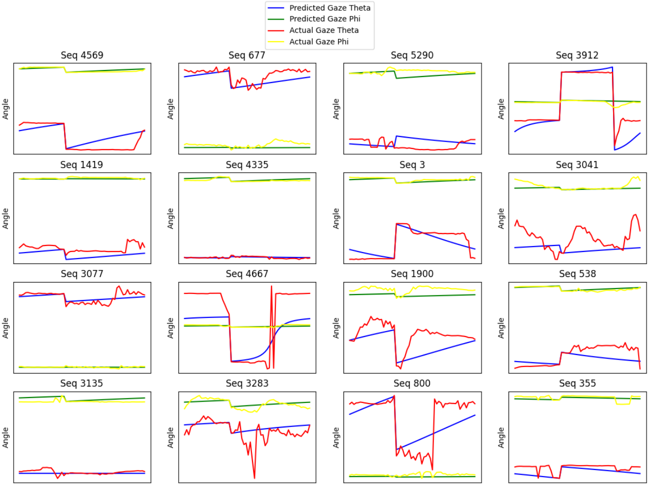
- A scipy 1D interpolation modal was used as a baseline to estimate the masked gaze values.
- A Single Modal Architecture which uses only gaze data for reconstruction was also used as a baseline.



Source: LIMU BERT Architecture

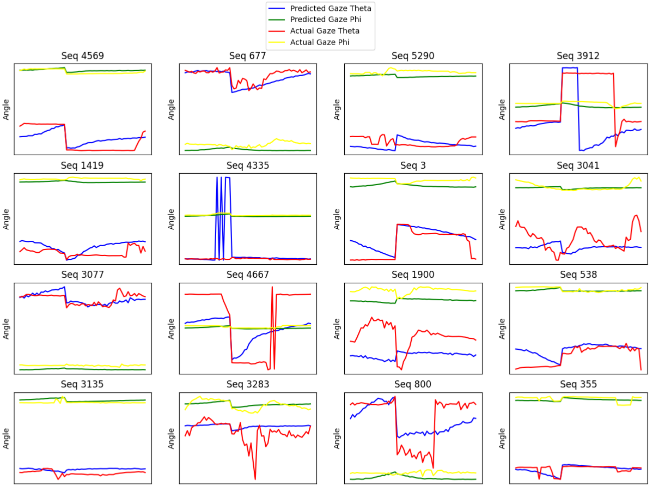
Table of Contents

Scipy Interpolation1D



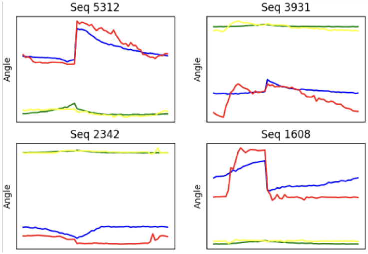
Source: Baseline results

Multi-Modal with Gaze Reconstruction

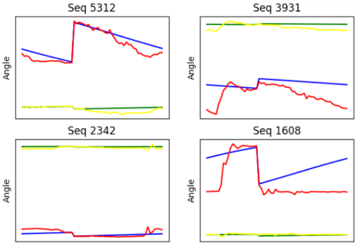


Source: Multi-Modal results

Gaze Reconstruction Comparison

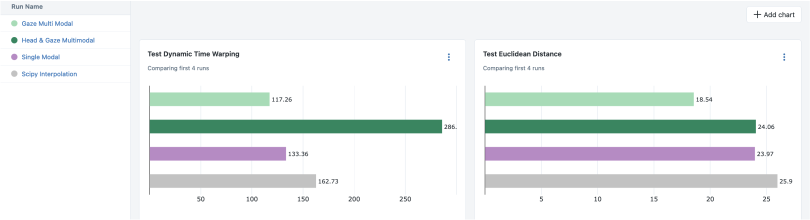


Source: Multi-Modal

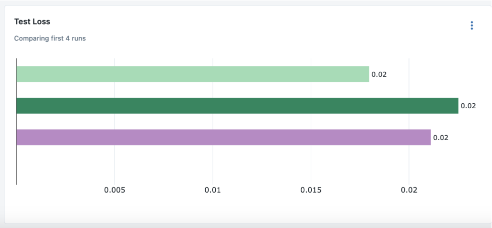


Source: Baseline

Summary



Source: Metrics



Source: Loss

- The Gaze Multimodal effectively reconstructs gaze values by leveraging the correlation between head and gaze data.
- The next step is to assess the model's performance on downstream applications.

Y. Jin, J. Liu, F. Wang, and S. Cui. Where are you looking?: A large-scale dataset of head and gaze behavior for 360-degree videos and a pilot study. In *MM '22: Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*. ACM, 2022. doi: 10.1145/3503161.3548200.

H. Xu, P. Zhou, R. Tan, M. Li, and G. Shen. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 220–233, 2021.