Unsupervised Gloss-Free Sign Language Recognition Using Transformer Model

Tamer Yeşildağ - Eray Taymaz

Advisor: Assoc. Prof. Hacer Yalım Keleş

Introduction

Sign language is a rich and complex visual language that plays a crucial role in communication for the Deaf and hard-of-hearing communities. However, building automatic systems to recognize sign language remains a difficult challenge due to limited annotated datasets and the diversity of signing styles.

In this project, we build upon the **Sign2GPT**[1] framework and focus on the **pretraining stage**, omitting the translation and language modeling components.

This project addresses these challenges by developing a **gloss-free, unsupervised sign language recognition model**. Unlike traditional systems that depend on gloss annotations which are textual labels describing individual signs we train our model using **pseudo-glosses**, which are automatically extracted keywords representing the meaning of sign sequences. This approach removes the need for costly manual annotation.

Our architecture leverages a **frozen DINOv2 Vision Transformer** for extracting spatial features from video frames and a **custom transformer encoder** to model temporal dynamics. The model is trained using a prototype-based contrastive loss that encourages the recognition of relevant sign components without relying on gloss order or sentence alignment.

By focusing on unsupervised learning and gloss-free recognition, this project demonstrates a scalable and efficient approach to sign language understanding that can adapt to real-world data with minimal supervision.

Dataset

For this project, we use the **RWTH-PHOENIX-Weather 2014T** [2] dataset, a benchmark dataset widely used in sign language recognition and translation research. It contains **German Sign Language (DGS)** recordings collected from real-world **weather forecast broadcasts**.

The dataset includes:

•**Video recordings** of professional signers delivering weather reports.

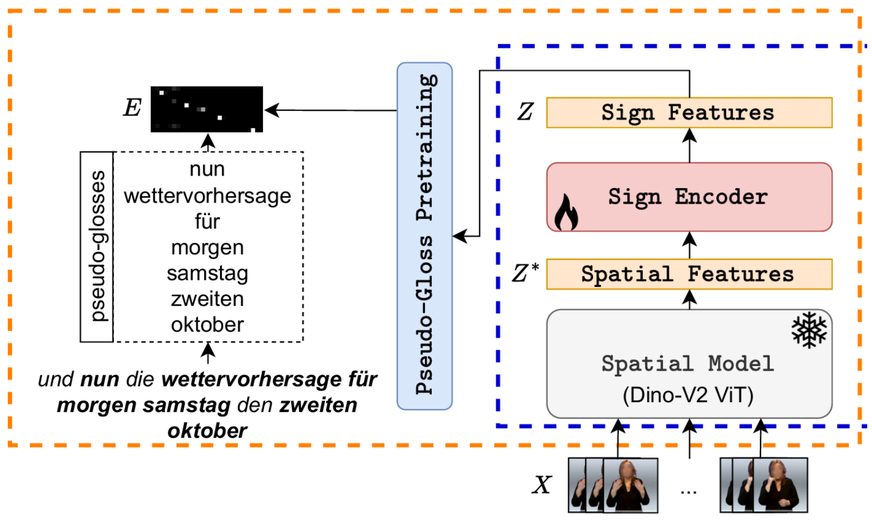
•**Aligned spoken language transcripts** in German for each video.

•**Train/validation/test splits**, with over **7,000 sentence-level video clips** in total.

Each video consists of a **continuous sequence of signs**, often several seconds long, recorded at a high frame rate. The dataset reflects real broadcast conditions, including varying signer poses, lighting, and speed making it ideal for evaluating models under realistic conditions.

In our gloss-free approach, we do **not use the manual gloss annotations** provided. Instead, we extract **pseudo-glosses** directly from the spoken German sentences, enabling the model to learn visual-semantic representations without relying on manually labeled gloss data.

Model Architecture



Feature Extraction:

- Raw video frames are first processed using a frozen DINOv2 Vision Transformer (ViT-S/14).
- DINOv2 extracts spatial embeddings from each video frame, resulting in a sequence of 384-dimensional feature vectors.
- These embeddings serve as the input to our transformer-based sign encoder.

Sign Transformer Encoder:

The core of the model is a custom Transformer encoder that learns temporal relationships in the sequence of embeddings. It includes:

- Positional Encoding: Adds sinusoidal positional encodings to retain temporal order information.
- Transformer Layers: A stack of 4 Transformer encoder layers with multi-head self-attention and feed-forward networks.

Prototype-Based Pretraining Module

- The final model learns to detect pseudo-glosses by comparing encoded features to a fixed set of 300-dimensional pseudo-gloss prototypes (based on fastText embeddings).
- Cosine similarity, softmax, and a custom localization matrix are used to predict the presence of each pseudo-gloss in the video.
- The model is trained using binary cross-entropy loss against pseudo-gloss presence labels.

References

- [1] R. Wong, N. C. Camgoz, and R. Bowden, “Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation,” *arXiv preprint arXiv:2405.04164*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2405.04164
- [2] H. Ney et al., “RWTH-PHOENIX-Weather 2014T: Sign language translation corpus,” RWTH Aachen University. [Online]. Available: https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-2014-T/

Results

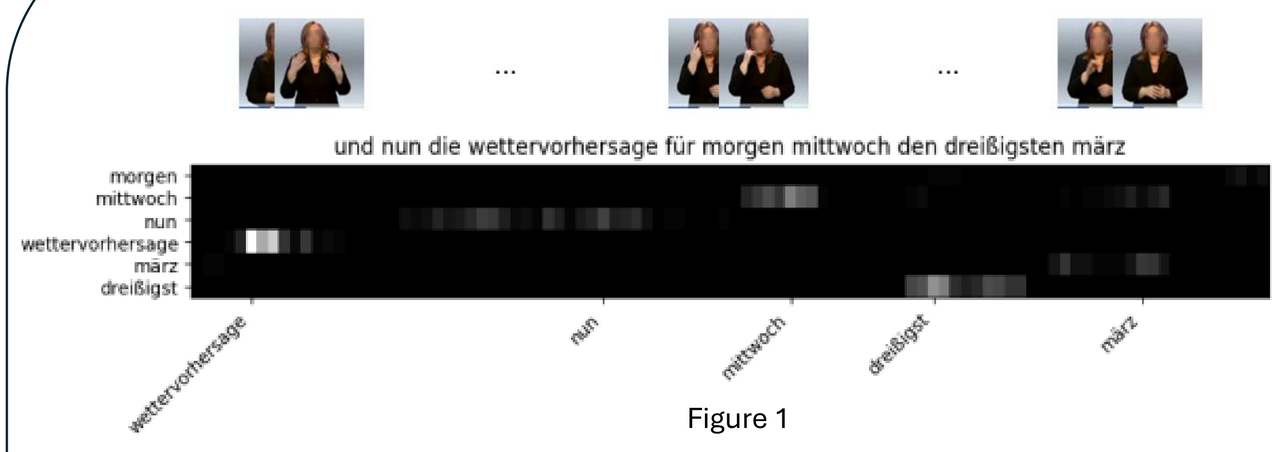


Figure 1

Figure 2

Left are the localization matrices (time × pseudo-gloss) for two test examples. Brighter regions indicate higher confidence that a given pseudo-gloss occurs at that point in the video.

Figure 1. “und nun die wettervorhersage für morgen mittwoch den dreißigsten märz” The model peaks on **wettervorhersage** at the very beginning, then on **nun**, and later on **mittwoch**, **dreißigsten**, and **märz** exactly when those signs are performed.

Figure 2. “der wind weht schwach bis mäßig aus süd bis südwest” Clear activations appear for **schwach**, **mäßig**, and **wind**, and a distinct, though lower-amplitude, response for **südwest**, demonstrating effective unsupervised temporal grounding of multiple glosses.