**Documentation**

**Overview**

This project builds a model to map glosses (word representations) to sentences using a combination of BERT-based gloss extraction, ResNet-BiLSTM visual feature extraction, cross-modal attention, compact bilinear pooling, and CTC loss.

**Model Pipeline**

The model consists of four major components:

1. Sentence Processing

Pretrained BERT (bert-large-uncased) extracts gloss features for selected words.

The extracted features represent words in a sentence, which will later be matched to corresponding visual features.

2. Visual Feature Extraction

ResNet-18 extracts spatial features from input frames.

The extracted features are passed through BiLSTM to capture temporal dependencies.

The output is stored as visual features.

3. Cross-Modal Attention

Multi-Head Attention (MHA) is used to align gloss features (BERT output) with visual features (ResNet-BiLSTM output).

A cross-modal loss is computed to ensure proper alignment.

The output consists of attended visual features.

4. Classification & CTC Loss

Compact Bilinear Pooling (Tensor Sketch) fuses gloss features and attended visual features.

A classifier is applied to predict glosses.

The final step applies Connectionist Temporal Classification (CTC), where the target is a sentence.