1 Spatial-temporal transformer for end-to-end sign language recognition

1.1 Introduction

Continuous sign language recognition (CSLR) is an essential task for communication between hearing-impaired and people without limitations, aiming to align low-density video sequences with high-density text sequences. This paper designed an end-to-end CSLR network: Spatial Temporal Transformer Network (STTN) to balance the spatial and temporal features during visual feature extraction. In the ST-encoder, part of the attention module focuses only on the contextual features on the temporal dimension and the other part extracts the spatial dynamic features of the video frames. To reduce the computational complexity of the large frame sequences, a patch operation is designed to map them into east-to-process sequences. A progressive learning strategy was used to explore the effectiveness of frame and patch size on the recognition results.

1.2 Spatial Temporal Transformer Network

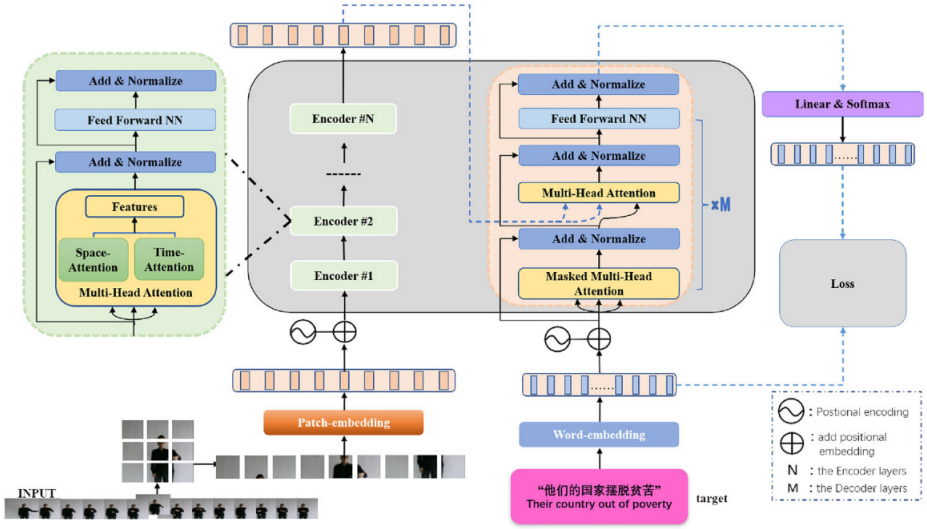

Figure 1: Spatial Temporal Transformer Network

At first patch embedding is performed. The input is of the vector sequence of size B * T * C * H * W where B is the batch size, T is no. of frames, C is no. of channels, and H and W represent the height and width of the frame. Each frame of size C * H * W is reshaped into a 2D block of dimension (h * w) * (p1 * p2 * C). The output of this patch embedding is B * T * N * D where N is the product of h * w.

To prevent the position-related loss in the network, the feature map with dimension B * T * N * D is position encoded.

Since CSLR is highly ST dependent, and it is difficult to capture temporal features as well as spatial features, an ST encoder structure for the dynamic spatial correlation and long-term temporal correlation of sign language videos was proposed. The incoming sign language video

vector is divided into two channels for processing temporal and spatial attention and then the extracted features are attached together. The decoder is the same as in the standard transformer architecture.