Key Action And Joint CTC-Attention Based Sign Language Recognition

Haibo Li   Liqing Gao   Ruize Han   Liang Wan   Wei Feng*

{lihb, lqgao, han_ruize, lwan, wfeng}@tju.edu.cn

College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China

Key Research Center for Surface Monitoring and Analysis of Cultural Relics, SACH, China

1. Introduction

Background and Motivation:

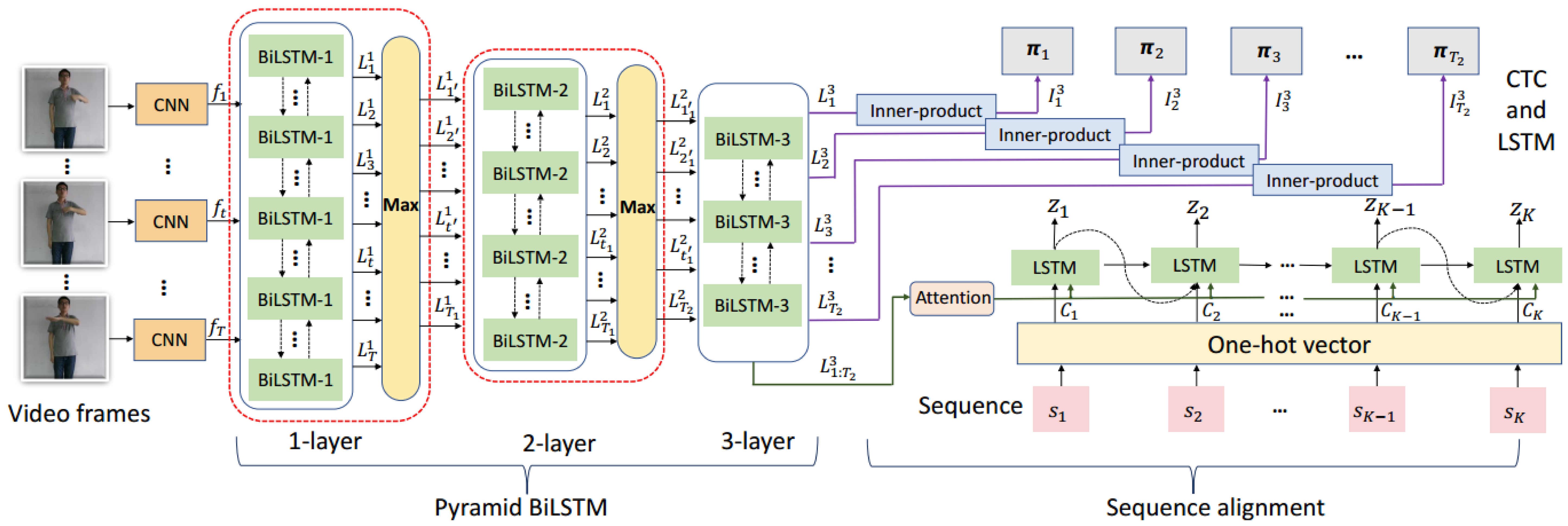
- Sign language video, owning a large number of redundant frames, is necessary to be selected the essential.
- Sign language video is characterized as continuous and dense action sequence, which is difficult to capture key actions corresponding to meaningful sentence.
- Connectionist Temporal Classification(CTC) based method assumes that the targets are conditionally independent, which can not capture context semantic.
- Encoder-Decoder based methods are sensitive to the data with noise, which can not handle the complex application very well.

We propose in this paper:

- A pyramid BiLSTM for video feature representation, which can also search the key actions over temproal scales.
- An LSTM to capture context semantic from target sentence and jointly train the framework using the CTC-attention based strategy.

2. Network Architecture

The architecture of our proposed method for Sign Language Recognition(SLR):



3. Loss Function

CTC-based loss function:

$$p(S|X) = \sum_{\pi \in \beta^{-1}(S)} p(\pi|X) \tag{1}$$

$$\mathcal{L}_{\mathrm{CTC}} = -\ln(p(S|X)) \tag{2}$$

LSTM-based loss function:

$$p(S|X) = \prod_{k=1}^{K} Z_{k,s_k} \tag{3}$$

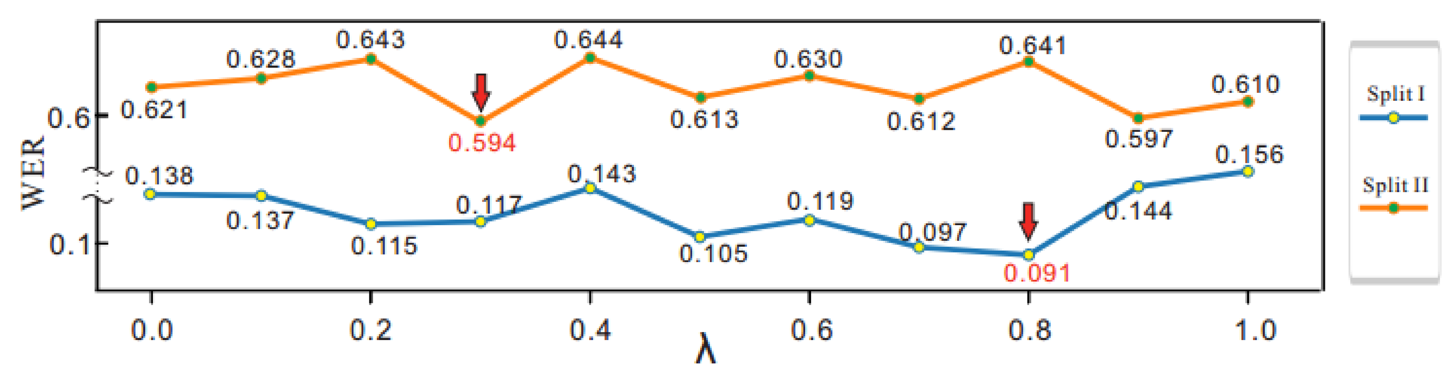
$$\mathcal{L}_{\mathrm{LSTM}} = -\ln(p(S|X)) \tag{4}$$

Total loss:

We use λ to weight the above the two loss functions in Eq.5

$$\mathcal{L} = \lambda \mathcal{L}_{\mathrm{CTC}} + (1-\lambda)\mathcal{L}_{\mathrm{LSTM}} \tag{5}$$

WER scores on CSL of proposed method using different λ:



4. Experiments Results

Table 1. Comparative results of different models.

Model	WER(%) ↓ Split I	WER(%) ↓ Split II
LSTM&CTC (Warp CTC)	15.6	63.1
S2VT[17]	29.8	62.5
LSTM-local-Attention [12]	18.9	62.7
LSTM-global-Attention [12]	12.1	62.1
DVWB[18]	13.7	61.7
Ours	**9.1**	**59.4**

Table 2. Ablation study of the proposed method.

Method	WER(%) ↓ Split I	WER(%) ↓ Split II	Method	WER(%) ↓ Split I	WER(%) ↓ Split II
SW-4/2	23.4	62.6	SW-4/4	13.7	64.5
SW-8/4	**9.1**	**59.4**	SW-8/8	13.4	65.2
w/o K	15.7	63.6	w/o CTC	13.8	62.1
w/o P	18.5	64.5	w/o LSTM	15.6	61.0
Last	15.7	63.6	Mean	15.4	63.0
Random	13.9	64.7	Ours	**9.1**	**59.4**

- The method we proposed worked better than existing one as shown in Table 1.
- Table 2 proves the effectiveness of the proposed method.

5. Conclusions

Conclusions:

- We proposed a pyramid BiLSTM to extract representations of key actions and capture the relation among them.
- We proposed to jointly train CTC and LSTM in order to integrate the advantages of both.