KEY ACTION AND JOINT CTC-ATTENTION BASED SIGN LANGUAGE RECOGNITION

Haibo Li, Liqing Gao, Ruize Han, Liang Wan, Wei Feng[†]

College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China
Key Research Center for Surface Monitoring and Analysis of Cultural Relics, SACH, China
{lihb, lqgao, han_ruize, lwan, wfeng}@tju.edu.cn

ABSTRACT

Sign Language Recognition (SLR) translates sign language video into natural language. In practice, sign language video, owning a large number of redundant frames, is necessary to be selected the essential. However, unlike common video that describes actions, sign language video is characterized as continuous and dense action sequence, which is difficult to capture key actions corresponding to meaningful sentence. In this paper, we propose to hierarchically search key actions by a pyramid BiLSTM. Specifically, we first construct three BiLSTMs to produce temporal relationships among input video sequence. Then, we associate these BiLSTMs by searching the salient responses in two groups of fixed-scale sliding window and capture key actions. Additionally, in order to balance the sequence alignment and dependency, we propose to jointly train Connectionist Temporal Classification (CTC) and Long Short-Term Memory (LSTM). Experimental results demonstrate the effectiveness of the proposed method.

Index Terms— Sign language recognition, Key action extraction, CTC and LSTM, Joint training

1. INTRODUCTION

Sign Language Recognition (SLR) translates sign language videos into natural languages for the sake of establishment of a communication channel between deaf-mute and hearing people. Unlike other video-based tasks such as video captioning [1, 2], video classification [3, 4], action recognition [5] etc., SLR seeks to explore relation between video sequence and language sequence.

Recently, many works are proposed to tackle the task of SLR [6, 7, 8, 9]. Their common pattern consists of two main stages, i.e., video representation and sequence alignment. For video representation, visual features are extracted by deep convnet then sent to recurrent neural network to capture temporal relations between frames. Many Long Short-Term Memory (LSTM) based methods are introduced to capture temporal dependency, such as SubUNets [8], recurrent convolutional neural networks [10] and hierarchical attention network [9]. For the alignment between input video sequence and target sentence sequence, two typical strategies are used, i.e., CTC sequence modeling method without any predefined alignments and RNN encoder-decoder framework. Connectionist Temporal Classification (CTC) [11] is a popular sequence learning algorithm, which aligns the input sequence and target sequence. CTC does not rely on a prior alignment between input and output sequences, but integrates over all possible alignments during the model training, e.g., Wang et al. [6] propose a connectionist temporal fusion method and Pu et al. [7] use iterative optimization with CTC for SLR. Another approach is based on the RNN encoder-decoder which was firstly proposed for machine translation [12]. This model transforms the input sequence of variable length into a fixed dimensional vector representation by the encoder process, then the decoder operation recovers the output sequence from this vector representation, e.g., Neural Machine Translation (NMT) for sign language translation [8] and hierarchical LSTM based on encoder-decoder framework [9].

Despite above progress, two problems still exist in SLR. First, the sign language video usually has many frames, some of which are redundant even harmful for SLR. However, due to the characters of continuity and denseness within sign language video, it is hard to extract the effective information. Second, 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, e.g., SLR. In this paper, inspired by multi-task learning in speech recognition [13, 14], we propose a pyramid BiLSTM to extract features for key actions from sign language videos. We further introduce an LSTM to capture context semantic from target sentence and jointly train the framework using the CTC-attention based strategy.

Our contributions are as follows: (1) We develop a pyramid BiLSTM for video feature representation, which can also extract the key actions over temporal scales. (2) We design to jointly train CTC and LSTM in order to capture the context semantic information while handle the complex application conditions. (3) Experimental results on dataset CSL demonstrate the effectiveness of the proposed method.

[†] is the corresponding author. This work was supported by NSFC 61671325 and 61572354.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

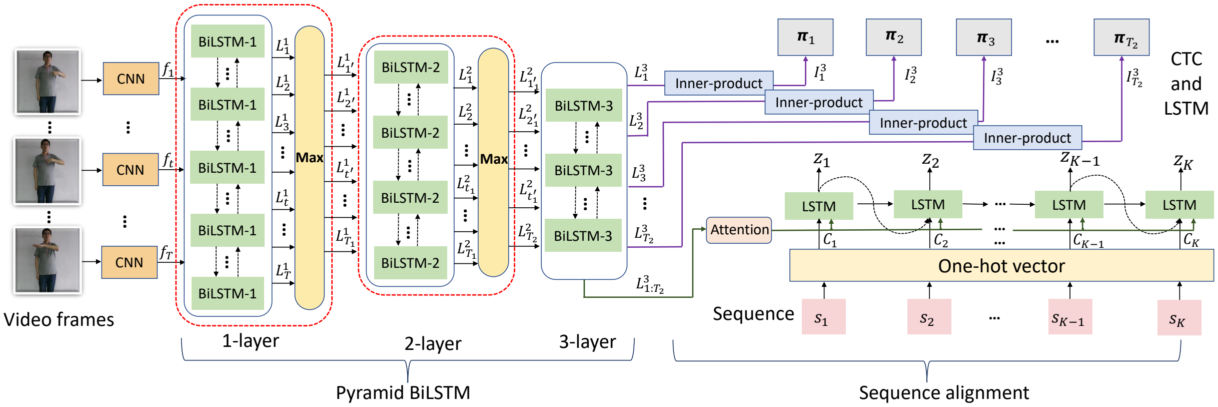

Fig. 1. The architecture of our proposed method for SLR: pyramid BiLSTM for video representation and joint training of CTC and LSTM. The pyramid BiLSTM contains three layers, 1-layer and 2-layer followed with maximum operation, respectively. The outputs of 3-layer BiLSTM are input of sequence alignment model.

2. THE METHOD

Fig. 1 illustrates the framework of our proposed method. We first use a basic CNN, e.g., ResNet-152 [15] to extract the spatial feature of each frame. Then a pyramid BiLSTM network is proposed to learn a high-level representation of the video, which aims to progressively get the hierarchical relationships from frames to an action, from actions to a word, from words to a sentence. The proposed network can well extract the hierarchical features for arbitrary-length videos. In addition, an LSTM network is presented for modeling the mutual relationships between the words in the target sentence. Finally, we use two losses in the whole network, i.e., the CTC loss and LSTM loss.

2.1. Hierarchical key action extraction

As shown in Fig. 1, we extract the image feature by removing the final fully connected layer in ResNet. Let a sign language video is represented as $X = (x_1, \cdots, x_T)$, where x_t represents the frame of a video and T denotes the number of frames in the current video. We use $\phi_\omega(\cdot)$ to represent the feature extraction by ResNet, then we obtain video features $F = (f_1, \cdots, f_T) = \{\phi_w(x_t)\}_{t=1}^T$, where $F \in \mathbb{R}^{T\times 2048}$.

In order to extract a high-level representation of key actions, we design a pyramid BiLSTM architecture. Inspired by [16], we propose a three-layers BiLSTM, from left to right, the length of BiLSTM for each layer decreases, as shown in Fig. 1. The 3-layer BiLSTM is the shortest, which represents a sign language video that consists of several words. The 2-layer BiLSTM represents a word that consists of many sign language actions. The 1-layer BiLSTM represents an action consists of some frames. We use a sliding window to obtain the maximum value and set the overlap rate of 50% between the sliding windows, which is similar to max-pooling in the temporal domain. Through this strategy, the key actions of each layer are extracted. For a long video, this structure can reduce redundant frames.

Given the sign language feature f_t, where $1 \leq t \leq T$, we input f_t into the 1-layer BiLSTM

$$L_t^1 = \text{BiLSTM}(f_t, \overrightarrow{h}_{t-1}, \overleftarrow{h}_{t+1}), \tag{1}$$

where $1 \leq t \leq T$ and L_t^1 represents outputs of the 1-layer BiLSTM. Because the BiLSTM will expand the feature dimension, we use a linear projection layer to maintain the dimension unchanged. We set the length of sliding window N_1 in 1-layer BiLSTM and N_2 in 2-layer BiLSTM, respectively. The sliding window slides over the L_t^1 and outputs a maximum value for each sliding window. That is, we calculate the value of each unit for BiLSTM and select the maximum value in the current sliding window. In addition, the overlap between two adjacent sliding windows is 50%. Therefore, the features of key actions are obtained effectively

$$L_{t'}^1 = \text{Max}(L_t^1, L_{t+1}^1, \cdots, L_{t+N_1-1}^1), \tag{2}$$

where $t = \frac{N_1}{2}\left(t' - 1\right) + 1$ and $1 \leq t' \leq T_1 = \lfloor \frac{T-N_1}{N_1/2} \rfloor + 1$. Similar to 1-layer BiLSTM, we can get outputs of 2-layer $L_{t_1}^2, 1 \leq t_1 \leq T_1$ and key actions of 2-layer $L_{t'_1}^2, 1 \leq t'_1 \leq T_2 = \lfloor \frac{T_1-N_2}{N_2/2} \rfloor + 1$. Finally, the features of key actions are input into the 3-layer BiLSTM and the outputs are represented as $L_{t_2}^3, 1 \leq t_2 \leq T_2$. In this paper, the output of inner-product layer following with 3-layer BiLSTM corresponds to the probability distribution of word labels.

2.2. Sequence alignment

Connectionist Temporal Classification (CTC) [11] is usually introduced to learn an alignment between the input sequence and target sequence in SLR. Let the sign word vocabulary is represented as ν, which contains a "blank" label (-). Denote the intermediate label path of the input sequence as $\pi = (\pi_1, \cdots, \pi_{T_2})$, where each word in π belongs to ν. Given the input sequence X, the probabilities $p(\pi|X)$ of π

$$p(\pi|X) = \prod_{t_2=1}^{T_2} p(\pi_{t_2}|X) = \prod_{t_2=1}^{T_2} I_{t_2}^3|_{\pi_{t_2}}, \tag{3}$$

where $I_{t_2}^3|_{\pi_{t_2}}$ denotes the log-probability of label π_{t_2} at $I_{t_2}^3$. Define a many-to-one map β, which has two options that remove all blanks and repeated labels from the paths (e.g. β(-cc-a-tt-) = β(c-a-t) = cat). Thus, given the sentence sequence $S = (s_1, s_2, \cdots, s_K)$, where K is the number of words, the conditional probability of S is calculated by summing up the probabilities of all corresponding paths

$$p(S|X) = \sum_{\pi \in \beta^{-1}(S)} p(\pi|X), \tag{4}$$

The CTC loss is defined as the negative log likelihood of the ground truth character sequence as

$$\mathcal{L}_{\text{CTC}} = -\ln(p(S|X)), \tag{5}$$

To efficiently compute the probability $p(S|X)$, the forward-backward algorithm [11] is applied.

However, CTC predicts the probability of each output on conditional independence. That is, previous predictions do not affect the subsequent prediction. Hence, we use an LSTM to establish the dependency between the predicted results. LSTM generates a corresponding sentence from the proposed pyramid BiLSTM. In the training stage, LSTM maximizes the log-likelihood of the target sentence given the hidden states and the previous words. In the inference stage, we choose a word with maximum probability until it outputs the finishing token $\langle$EOS$\rangle$.

In LSTM network, we apply the attention mechanism. We define the vocabulary as $\nu^{'}$, which contains beginning token $\langle$SOS$\rangle$ and ending token $\langle$EOS$\rangle$. The output of LSTM is

$$h_k = \text{LSTM}(C_k, s_k, h_{k-1}), \tag{6}$$

where C_k is context vector relative with k-th word in sentence sequence $S = (s_1, s_2, \cdots, s_K)$, s_k is k-th word in S which will be encoded as one-hot vector for LSTM and h_{k-1} is a hidden state. We add a linear projection layer after LSTM to obtain categorical probability Z_k, the active value of word s in Z_k is represented as $Z_{k,s}$. given input video sequence X and sentence sequence S, the probability of S is defined as

$$p(S|X) = \prod_{k=1}^{K} Z_{k,s_k}, \tag{7}$$

The loss function of the LSTM is computed from

$$\mathcal{L}_{\text{LSTM}} = -\ln(p(S|X)), \tag{8}$$

In order to achieve joint training between CTC and LSTM, we use λ to weight the above two loss functions

$$\mathcal{L} = \lambda\mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{LSTM}}, \tag{9}$$

where $0 \le \lambda \le 1$, we can adjust λ to achieve effective results for SLR. Through the joint training strategy, on the basis of CTC prediction performance, LSTM is used to establish the dependency between sentence words and extract the context semantics of sentences.

3. EXPERIMENTAL RESULTS

3.1. Setup

Dataset and metrics. We use the popular SLR benchmark, i.e., Chinese Sign Language dataset (CSL)[1] to evaluate the proposed method, which contains 25K labelled videos by 50 singers and the vocabulary size is 178. We split the dataset with two strategies. 1) **Split I - singer independent test:** It splits the videos of 40 singers as the training set (20K videos) and that of the remaining 10 singers as testing set (5K videos). The sentences of training and testing sets are the same, but the singers are different. 2) **Split II - unseen sentences test:** We select 6 sentences as a testing set (1.5K videos), and the left 94 sentences as the training set (23.5K videos). Although some of the words in 6 testing sentences separately appear in the remaining 94 training sentences, the context of each word is completely different in occurrence order and application scenarios. We use the standard metric – Word Error Rate (WER) to evaluate the similarity between two sentences. Specifically, WER $= \frac{\text{num\_ins}+\text{num\_del}+\text{num\_sub}}{\text{num\_words}} \times 100\%$, where num_ins, num_del and num_sub denote the number of insertion, deletion and substitution operations to transform predicted sentence into the ground truth. The num_words represents the number of words in the ground truth. The smaller WER means the better recognition.

Implementation details. We use the pretrained Resnet-152 for feature extraction. The sliding window size used in the 1-layer and 2-layer of the pyramid BiLSTM is set as 8 and 4, respectively. We use the Adam algorithm for loss optimization. The initial learning rate and weight decay are set to 1×10^{-3} and 5×10^{-5}, and we adopt Step-LR schedule to change the learning rate for every few epochs. The hidden states of the pyramid BiLSTM and LSTM are also set to 256. In order to set an optimal parameter λ in Eq. (9), we conduct experiments with different λ. As shown in Fig. 2, we can see that $\lambda = 0.8/0.3$ gets the best performance on Split I / II. Hence, λ is set to 0.8 / 0.3 on Split I / II in the following experiments.

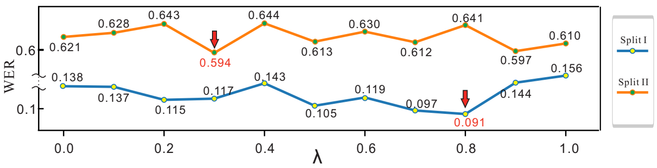

Fig. 2. WER scores on CSL of our method using different λ.

[1] http://mccipc.ustc.edu.cn/mediawiki/index.php/SLR_Dataset

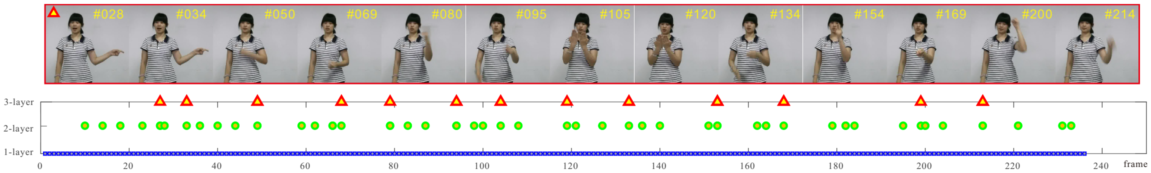

Fig. 3. Illustration of the hierarchical key action searching strategy.

3.2. Comparative results on CSL

In this section, we compare our method with some existing algorithms on the CSL dataset using WER metric. For a fair comparison, we use the same features extracted by pretrained RestNet model for all the comparative methods. As shown in Table 1, **LSTM&CTC** model is widely used in sequential data analysis, e.g., speech recognition. The WER scores on Split I and II of **LSTM&CTC** are 15.6% and 63.1%, respectively. Meanwhile, we also compare our model to some existing encoder-decoder models. **S2VT** [17] uses a standard two-layers stacked LSTM architecture with fixed encoder length, which achieves WER scores of 29.8% and 62.5% on Split I and II, respectively. **LSTM-local-Attention** [12] and **LSTM-global-Attention** [12] use different attention mechanisms in LSTM to learn the alignments between input video sequence and output target sequence, which achieve WERs of 18.9% / 62.7%, 12.1% / 62.1% on Split I / II, respectively. **DVWB** [18] integrates BiLSTM and a soft attention mechanism to generate better global representations for videos, which achieves WERs of 13.7% and 61.7% on Split I and II, respectively. We can see that the proposed method outperforms all other competitors in both Split I and II and achieves the WER scores of 9.1% and 59.4%.

Table 1. Comparative results of different models.

Model	WER(%) $\downarrow$	
	Split I	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**

3.3. Ablation study

Then, we validate the effectiveness of each component of our method. As shown in the first two rows of Table 2, we first change the length of the sliding window in key action selection stage discussed in Section 2.1, e.g., 'SW-4/2' denotes our method using the sliding windows length of $N_1 = 4$ and $N_2 = 2$ in 1-layer and 2-layer, respectively. We can see that our method with the setting $N_1 = 8$ and $N_2 = 4$ provides the best performance in both Split I and II. Next, we verify the effectiveness of the pyramid BiLSTM network and sequence alignment strategy, which is shown in the middle two rows in Table 2. We can see that the SLR WER of the proposed method without key action selection strategy, i.e., 'w/o K' increases into 15.7% and 63.6% on Split I and II, respectively. We then remove the pyramid BiLSTM network, which is denoted as 'w/o P'. We can also see that the WER further increases compared to 'w/o K'. To validate the impact of two losses for sequence alignment described in Section 2.2, we remove the branch using CTC loss and LSTM loss, respectively. We can see that the proposed method using single loss, either CTC or LSTM, provides worse performance than using two. The comparison results demonstrate that CTC loss and LSTM loss can be regarded as a pair of complementary losses to help each other in our task. We further study the validity of the key action selection method in Section 2.1. We consider three alternative methods: in each sliding window 1) we simply select the *last* hidden state of output feature 2) we calculate the *mean* value of all the output feature 3) we *randomly* select one from all the output features. From the last two rows in Table 2, we can see that the proposed key action selection approach by computing the maximum of the feature confidence score is superior than all other three methods.

Table 2. Ablation study of the proposed method.

Method	WER(%) $\downarrow$		Method	WER(%) $\downarrow$	
	Split I	Split II		Split I	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**

We finally illustrate the key action searching strategy in Fig. 3. The bottom figure shows the selective frame index in each layer. We further illustrate the key frames generated by the last layer in the top figure. We find that the selective frames include all the key actions of the whole video. Note that, a sign language volunteer can obtain the meaning of the whole sequence only by the selective frames.

4. CONCLUSION

In this paper, we have proposed a pyramid BiLSTM to extract representations of key actions and capture the relation among them. Specifically, we construct the pyramid BiLSTMs to produce the hierarchical relationships from video frames to target sentence, in which we extract the key actions for calculation cost reducing and effective information capturing. Besides, we also propose to jointly train CTC and LSTM in order to integrate the advantages of both. Experimental results on CSL benchmark demonstrate that our proposed method outperforms the state-of-the-art methods.

5. REFERENCES

[1] S. Venugopalan, H.J. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, “Translating videos to natural language using deep recurrent neural networks,” *arXiv preprint arXiv:1412.4729*, 2014.

[2] Z.Q. Shen, J.G. Li, Z. Su, M.J. Li, Y.R. Chen, Y.G. Jiang, and X.Y. Xue, “Weakly supervised dense video captioning,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[3] K. Andrej, T. George, S. Sanketh, L. Thomas, S. Rahul, and F.F. Li, “Large-scale video classification with convolutional neural networks,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[4] J.H. Yoo, “Large-scale video classification guided by batch normalized (lstm) translator,” *arXiv preprint arXiv:1707.04045*, 2017.

[5] L.M. Wang, Y.J. Xiong, Z. Wang, Y. Qiao, D.H. Lin, X.O. Tang, and V.G. Luc, “Temporal segment networks: Towards good practices for deep action recognition,” in *Proceedings of the European Conference on Computer Vision*, 2016.

[6] S. Wang, D. Guo, W.G. Zhou, Z.J. Zha, and M. Wang, “Connectionist temporal fusion for sign language translation,” in *ACM Multimedia Conference on Multimedia Conference*, 2018.

[7] J.F. Pu, W.G. Zhou, and H.Q. Li, “Dilated convolutional network with iterative optimization for continuous sign language recognition.,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

[8] N.C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, “Subunets: End-to-end hand shape and continuous sign language recognition,” in *Proceedings of IEEE International Conference on Computer Vision*, 2017.

[9] D. Guo, W.G. Zhou, H.Q. Li, and M. Wang, “Hierarchical (lstm) for sign language translation,” in *AAAI Conference on Artificial Intelligence*, 2018.

[10] R.P. Cui, H. Liu, and C.S. Zhang, “Recurrent convolutional neural networks for continuous sign language recognition by staged optimization,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[11] A. Graves, S. Fernández, F. Gomez, and J.ürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of International Conference on Machine Learning*, 2006.

[12] M.T. Luong, H. Pham, and C.D. Manning, “Effective approaches to attention-based neural machine translation,” *arXiv preprint arXiv:1508.04025*, 2015.

[13] S. Kim, T. Hori, and S. Watanabe, “Joint ctc-attention based end-to-end speech recognition using multi-task learning,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[14] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, “Advances in joint ctc-attention based end-to-end speech recognition with a deep (cnn) encoder and (rnn-lm),” *arXiv preprint arXiv:1706.02737*, 2017.

[15] K.M. He, X.Y. Zhang, S.Q. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[16] J.S. Su, J.L. Zeng, D.Y. Xiong, Y. Liu, M.X. Wang, and J. Xie, “A hierarchy-to-sequence attentional neural machine translation model,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 3, pp. 623–632, 2018.

[17] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to sequence-video to text,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[18] Y. Bin, Y. Yang, F.M. Shen, N. Xie, H.T. Shen, and X.L. Li, “Describing video with attention-based bidirectional lstm,” *IEEE Transactions on Cybernetics*, vol. 49, no. 7, pp. 2631–2641, 2018.