Dense Temporal Convolution Network for Sign Language Translation

Dan Guo[1], ShuoWang[1], Qi Tian[2], Meng Wang[1]

[1]*School of Computer Science and Information Engineering, Hefei University of Technology*

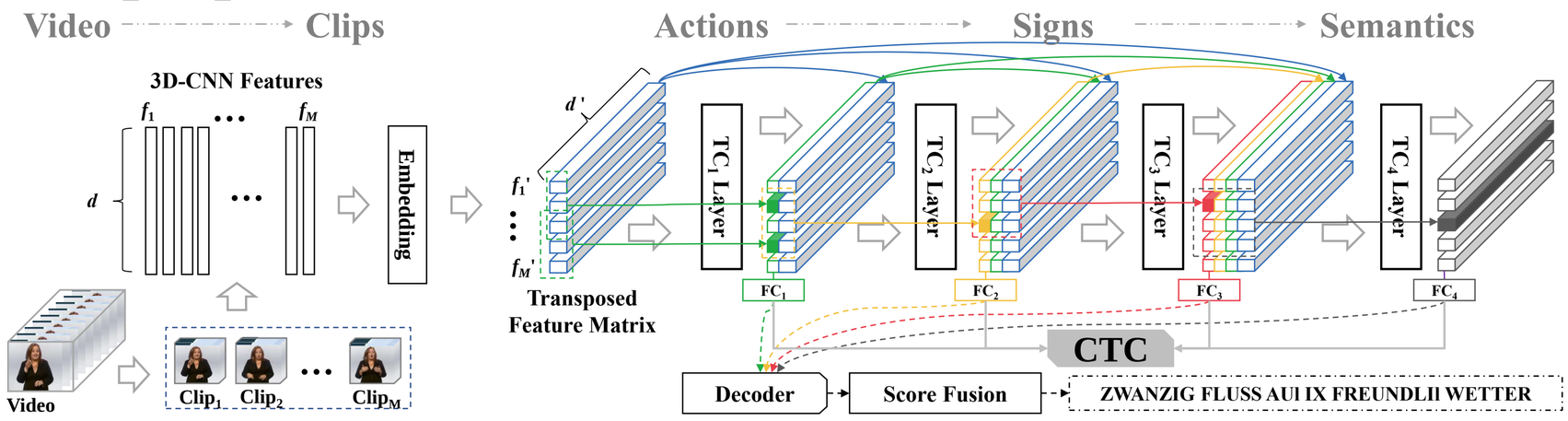
[2]*Huawei Noah's Ark Lab Department of Computer Science, University of Texas at San Antonio*

guodan@hfut.edu.cn, shuowang.hfut@gmail.com, tian.qi1@huawei.com, eric.mengwang@gmail.com

Abstract

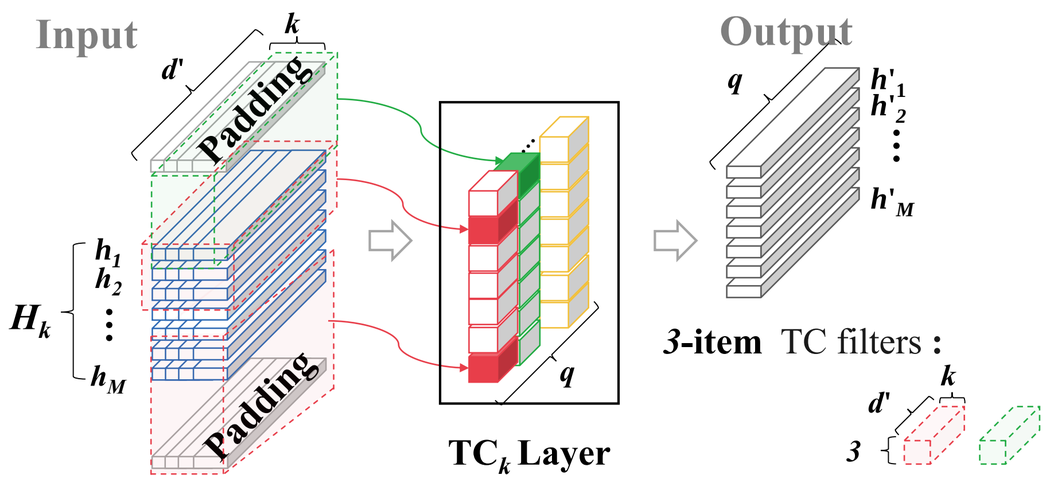
We propose a dense temporal convolution network, termed DenseTCN which captures the actions in hierarchical views. Within this network, a temporal convolution (TC) is designed to learn the short-term correlation among adjacent features and further extended to a dense hierarchical structure. In the k^{th} TC layer, we integrate the outputs of all preceding layers together: (1) The TC in a deeper layer essentially has larger receptive fields, which captures long-term temporal context by the hierarchical content transition. (2) The integration addresses the SLT problem by different views, including embedded short-term and extended long-term sequential learning. Finally, we adopt the CTC loss and a fusion strategy to learn the feature-wise classification and generate the translated sentence. The experimental results on two popular sign language benchmarks, i.e. PHOENIX and USTC-ConSents, demonstrate the effectiveness of our proposed method in terms of various measurements.

Overview of the proposed DenseTCN



Details: We first split the video into clips and extract the feature of each clip from the 3D-CNN, which captures the sequential and the spatial information simultaneously. Then, the multi-layered TC structure is developed for calculating adjacent features in different receptive fields. Meanwhile, we concatenate the outputs of all preceding layers and use them as the input of the current calculation layer. In the training stage, we use the CTC to learn the relationship between the translated and the real sentences in each TC layer. In the testing stage, the greedy decoder and the fusion strategy are used to find a more reliable sentence.

Dense TC



The operations of the TC_k layer (i.e. k = 4, n = 3). We consider a feature matrix contains M temporal features in d' dimension as input $H_k = \{h_i\}_{i=1}^{M} \in \mathbb{R}^{k \times M \times d'}$. Such matrix is concatenated of the outputs from the 0^{th} to $(k-1)^{th}$ calculation layer, we first pad it across the temporal dimension. Then we employ q TC filters to capture the dynamic visual information from the input by calculating n-item adjacent features. At last, we concatenate the outputs after all filters across the feature dimension into a matrix $\{h'_i\}_{i=1}^{M} \in \mathbb{R}^{M \times q}$ as the output of the k^{th} TC layer.

Sentence Learning

$$p_k = FC_k(O_k) = O_k \cdot W_k + b_k \,, k>0$$

$$Pr(\pi_k|p_k) = \prod_{j=1}^{M} Pr(\pi_{k,j}|p_k) \,, \forall \pi_{k,j} \in V_{OC'}$$

$$Pr(y|p_k) = \sum_{\pi_k \in B^{-1}(y)} Pr(\pi_k|p_k)$$

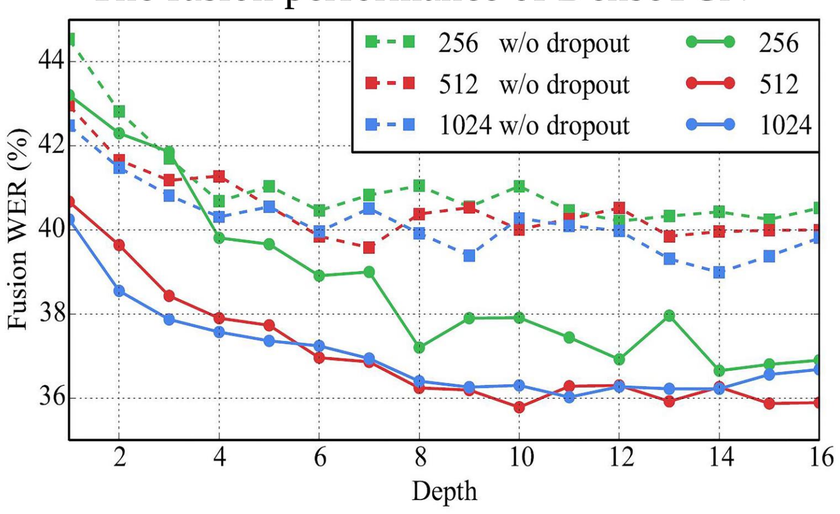
$$L_{CTC} = -log\, Pr(y|P) = -\sum_{k=1}^{K} log Pr(y|p_k)$$

Score Fusion and Translation

$$p^i_{fusion,j} = \frac{1}{K} \sum_{k=1}^{K} \frac{e^{p^i_{h,j}}}{\sum_{j'=1}^{w} e^{p^i_{k,j'}}}$$

Experiments

The fusion performance of DenseTCN



Evaluations under PHOENIX
(▲: Other modality, △: Extra supervision)

Methods	VAL		TEST	
	del / ins	WER	*del / ins*	WER
HOG-3D ▲	25.8 / 4.2	60.9	23.2 / 4.1	58.1
CMLLR ▲	21.8 / 3.9	55.0	20.3 / 4.5	53.0
1M-Hands ▲ △	16.3 / 4.6	47.1	15.2 / 4.6	45.1
CNN-Hybrid ▲ △	12.6 / 5.1	38.3	11.1 / 5.7	38.8
Staged-Opt ▲ △	13.7 / 7.3	39.4	12.2 / 7.5	38.7
SubuNets ▲	14.6 / 4.0	40.8	14.3 / 4.0	40.7
Dilated-CNN △	8.3 / 4.8	38.0	7.6 / 4.8	37.3
LS-HAN	-	-	-	38.3
CTF-SLT	12.8 / 5.2	37.9	11.9 / 5.6	37.8
DenseNet*	-	49.7	-	49.2
Our DenseTCN	10.7 / 5.1	**35.9**	10.5 / 5.5	**36.5**

Evaluation under USTC-ConSents

Methods	WER
DTW-HMM [Zhang *et al.*, 2014]	28.4
LSTM [Venugopalan *et al.*, 2015b]	26.4
S2VT [Venugopalan *et al.*, 2015a]	25.5
LSTM-A [Yao *et al.*, 2015]	24.3
LSTM-E [Pan *et al.*, 2016]	23.2
HAN [Yang *et al.*, 2016]	20.7
LS-HAN [Huang *et al.*, 2018]	17.3
HLSTM-atten [Guo *et al.*, 2018]	**10.2**
CTF-SLT [Wang *et al.*, 2018]	11.2
DenseNet* [Huang *et al.*, 2017]	38.3
Our DenseTCN	14.3

Methods	WER
S2VT [Venugopalan *et al.*, 2015a]	67.0
S2VT(3-layer) [Yao *et al.*, 2015]	65.2
HLSTM (SYS sampling) [Guo *et al.*, 2018]	66.3
HLSTM [Guo *et al.*, 2018]	66.2
HLSTM-atten [Guo *et al.*, 2018]	64.1
DenseNet* [Huang *et al.*, 2017]	52.1
Our DenseTCN	**44.7**

Acknowledgments: This work is supported by the National Natural Science Foundation of China (NSFC) under grants 61725203, 61732008, and 61876058.