Connectionist Temporal Modeling of Video and Language: a Joint Model for Translation and Sign Labeling

Dan Guo, Shengeng Tang, Meng Wang

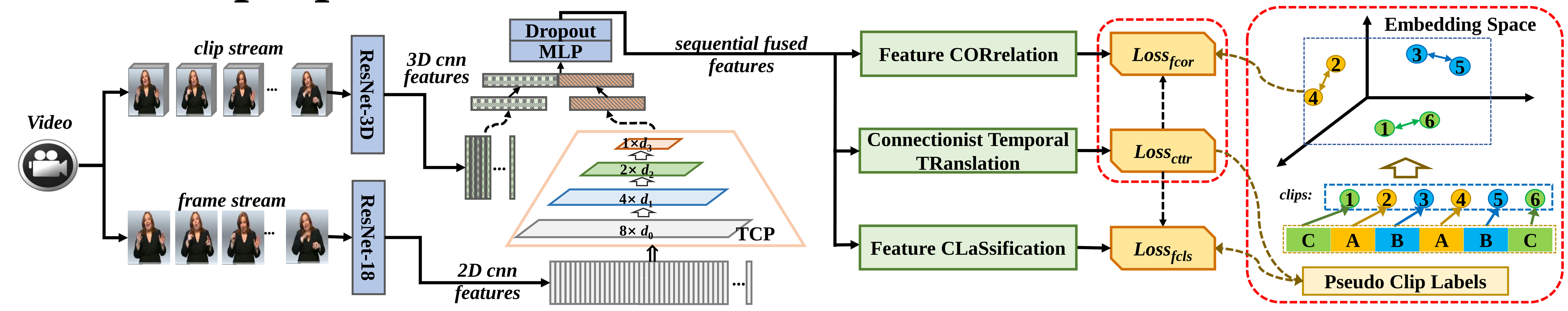
School of Computer Science and Information Engineering, Hefei University of Technology

guodan@hfut.edu.cn, tsg1995@hfut.edu.cn, eric.mengwang@gmail.com

Abstract

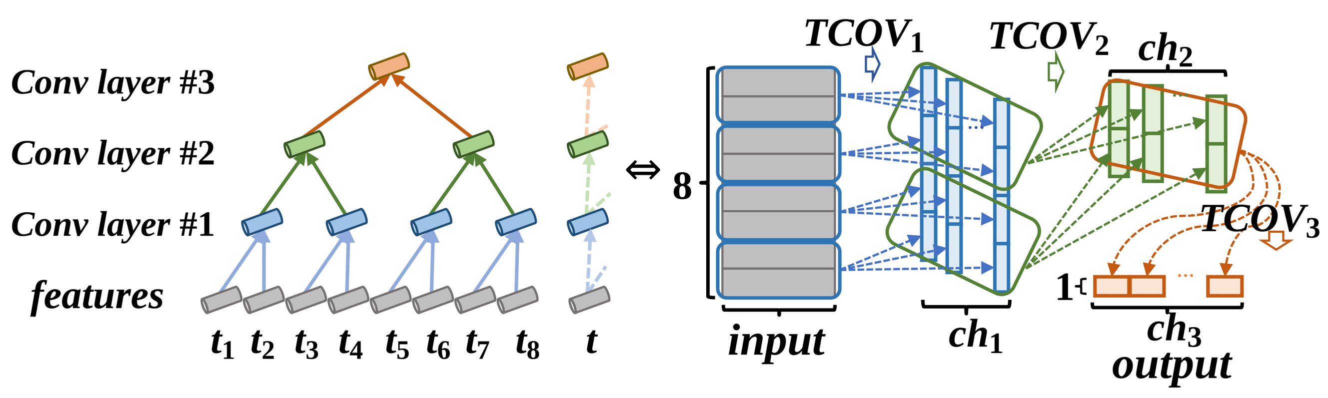
We propose a Connectionist Temporal Modeling (CTM) network for sentence translation and sign labeling. To acquire short-term temporal correlations, a Temporal Convolution Pyramid (TCP) module is performed to convert 2D CNN features to pseudo 3D' features. CTM aligns the pseudo 3D' with the original 3D CNN clip features and fuses them. Next, we implement a connectionist decoding scheme for long-term sequential learning. Here, we embed dynamic programming into the decoding scheme, which learns temporal mapping among features, sign labels, and the generated sentence directly. The solution using dynamic programming to sign labeling is considered as pseudo labels. Finally, we utilize the pseudo supervision cues in an end-to-end framework. A joint objective function is designed to measure feature correlation, entropy regularization on sign labeling, and probability maximization on sentence decoding. The experimental results using the RWTH-PHOENIX-Weather and USTC-CSL datasets demonstrate the effectiveness of the proposed approach.

Overview of the proposed CTM



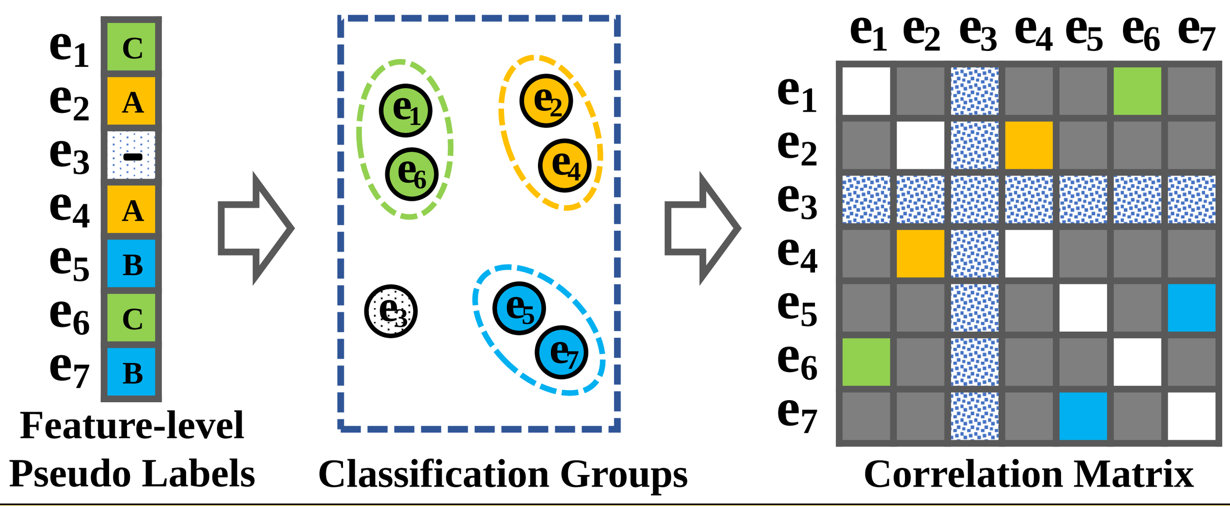
Details: Overview of the proposed CTM framework for online SLT. Given a video, we extract 2D frame-level and 3D clip-level feature streams using the pre-trained models ResNet-18 and ResNet-3D, respectively. The TCP module is conducted on the 2D features to learn short-term temporal clues, and align them to the 3D features. Then, the fused features are fed into three modules for long-term sequential learning. Finally, we utilize pseudo supervision cues in the online deep model. A joint loss optimization combining L_{fcor}, L_{cttr}, and L_{fcls}, is designed to measure feature correlation, sentence decoding, and entropy regularization on sign labeling.

TCP Module



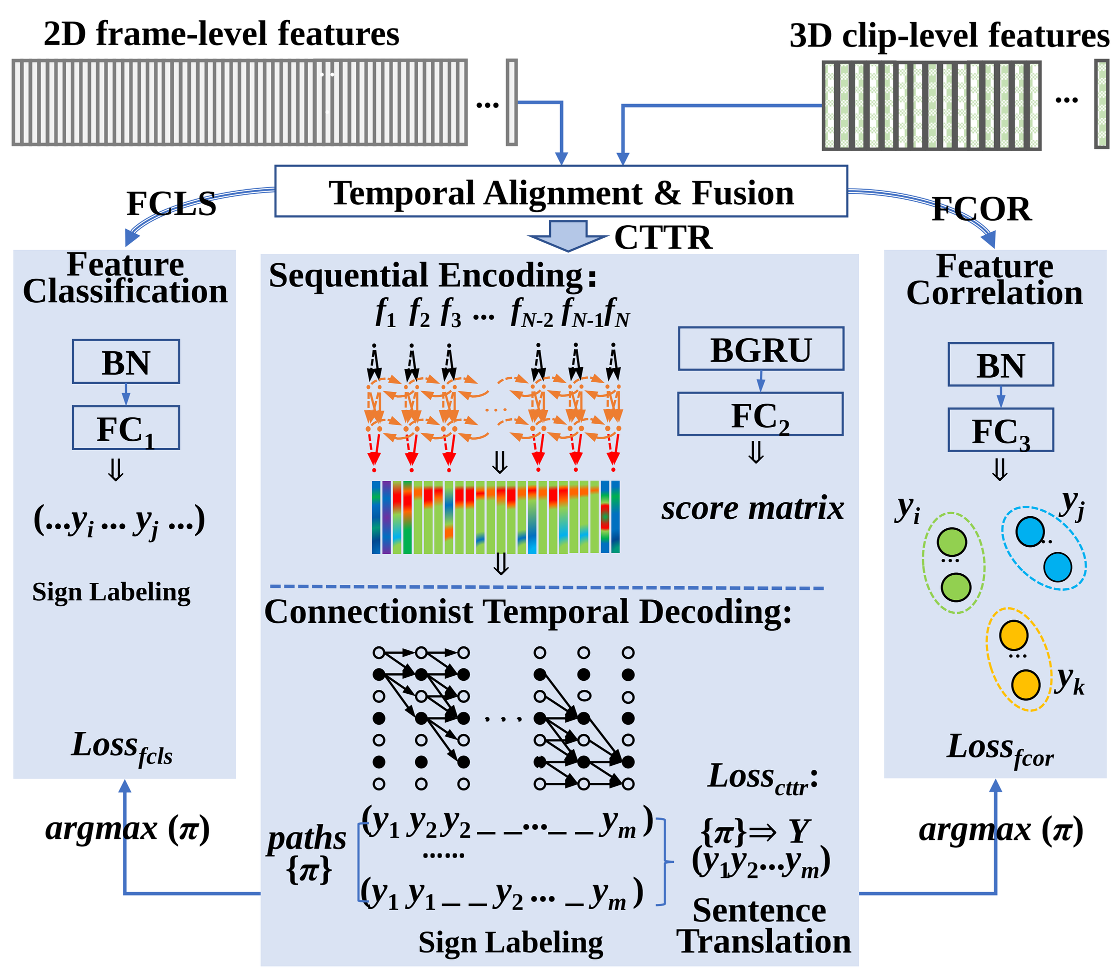
Temporal Convolution Pyramid (TCP) on 2D features.

Triplet Loss Calculation



Triplet loss calculation based on different classification groups for feature correlation. e_3 indicates a blank symbol '-'. In the matrix, we do not consider diagonals and squares with snowflakes, where self-correlation and the blank label '-' have no word meaning.

Online SLT



Architecture of the proposed approach for online SLT, which consists of a Connectionist Temporal Translation module (CTTR), a Feature Classification module (FCLS), and a Feature Correlation module (FCOR). The middle CTTR module decodes the connectionist mapping among features, words, and the generated sentence. Pseudo supervision cue π is utilized on both two side modules (FCLS and FCOR).

Joint Loss Optimization

$$L=\frac{1}{|S|}\sum L_{cttr}+\frac{1}{|M|}\sum L_{fcls}+\frac{1}{|T|}\sum L_{fcor}$$

Experiments

Performance Comparison on PHOENIX Dataset

Methods	Off-line Iterations	Other Modality	VAL(%)		TEST(%)	
			des/ins	WER	des/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
1-Mio-H	3	√	19.1/4.1	51.6	17.5/4.5	50.2
1-Mio-H+CMLLR	3	√	16.3/4.6	47.1	15.2/4.6	45.1
CNN-Hybrid	3	√	12.6/5.1	38.3	11.1/5.7	38.8
Staged-Opt-init	-	√	16.3/6.7	46.2	15.1/7.4	46.9
Staged-Opt	3	√	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-init	-		18.5/2.6	60.3	18.1/2.8	59.7
Dilated-CNN	5		8.3/4.8	38.0	7.6/4.8	37.3
Our Method	-		11.6/6.3	38.9	10.9/6.4	38.7

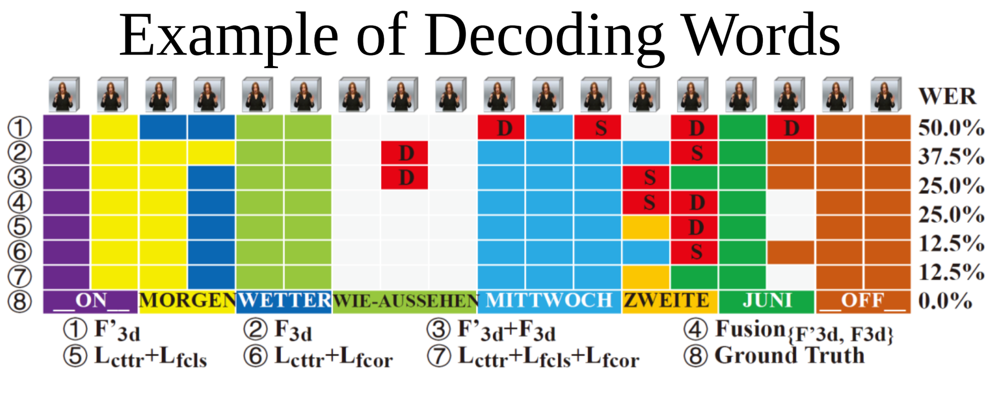
Performance with Different Features

Features	VAL(%)		TEST(%)	
	des/ins	WER	des/ins	WER
f_{2d}	55.1/1.5	69.4	53.6/1.8	58.1
f'_{3d}	27.5/5.8	63.6	26.8/6.1	53.0
f_{3d}	21.0/5.1	45.1	20.0/5.5	50.2
$f'_{3d} + f_{3d}$	10.5/7.3	42.2	10.8/7.8	45.1
Fusion$\{f'_{3d}, f_{3d}\}$	10.6/6.9	41.0	10.1/7.9	41.3

Performance with Different Loss

Loss	VAL(%)		TEST(%)	
	des/ins	WER	des/ins	WER
L_{cttr}	10.6/6.9	41.0	10.1/7.9	41.3
$L_{cttr}+L_{fcls}$	10.2/6.7	39.9	10.3/7.7	40.2
$L_{cttr}+L_{fcor}$	11.3/6.7	39.8	10.9/6.9	40.0
$L_{cttr}+L_{fcls}+L_{fcor}$	11.8/5.9	38.9	10.6/6.1	38.7

Example of Decoding Words



Performance comparison on USTC-CSL

Methods	TEST WER (%)
S2VT	58.1
S2VT(3-layer)	53.0
HLSTM	50.2
HLSTM-attn	45.1
Our Method	41.3

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