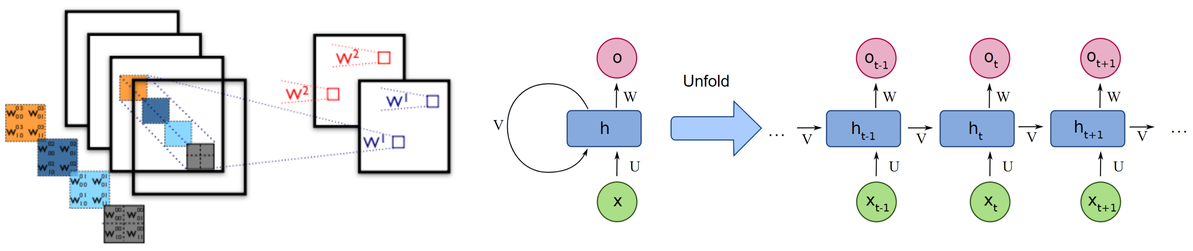
A Self-Attention Network for Skeleton-based Human Action Recognition (Poster #21)

Computational Imaging Lab. (CIL)
Department of Computer Science

Sangwoo Cho, Muhammad Maqbool, Fei Liu, Hassan Foroosh
{swcho, hasanmaqbool}@knights.ucf.edu, {feiliu, foroosh}@cs.ucf.edu
Dept. of Computer Science, University of Central Florida

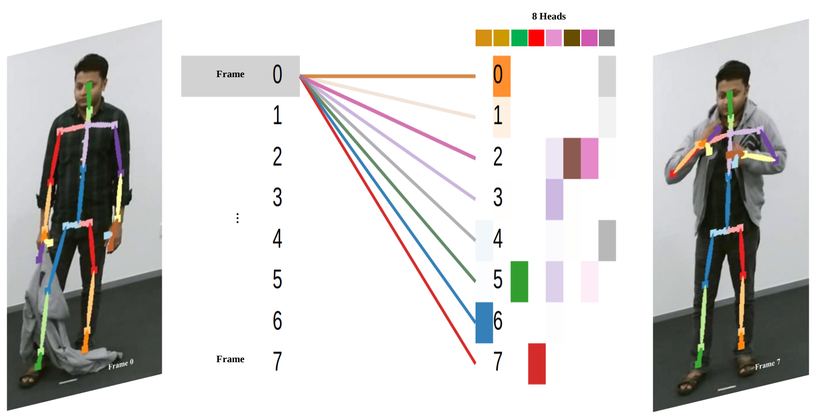
WACV 2020
March 1-5, Aspen, CO

Motivation



CNN (GCN) **LSTM**

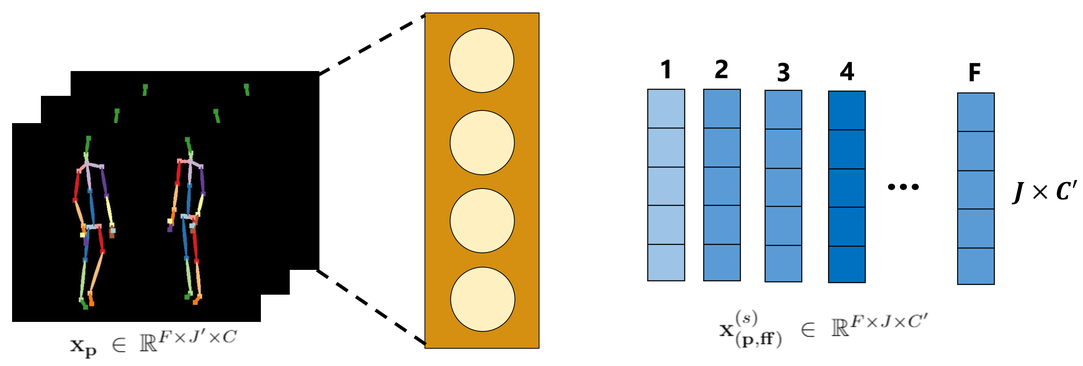
Local activation ➔ long-term information



Q) How to correlate joints in one frame to other frames?

Encoders

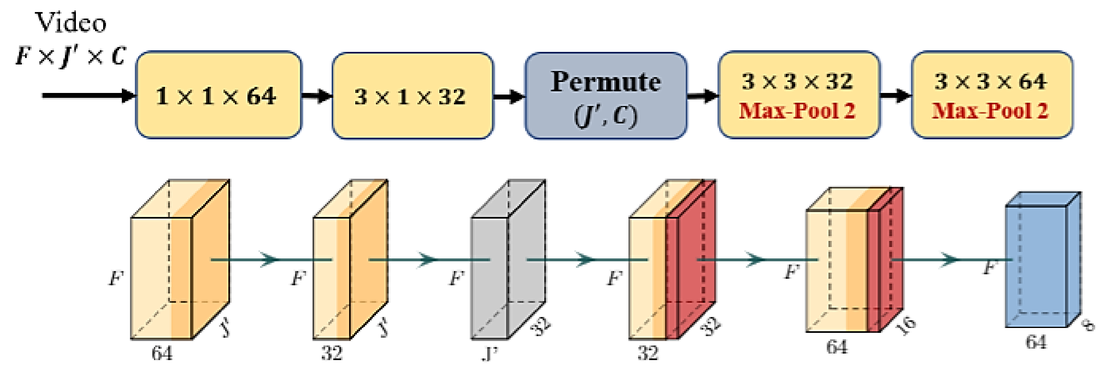
FF: Convert joint-based vector to higher dim.



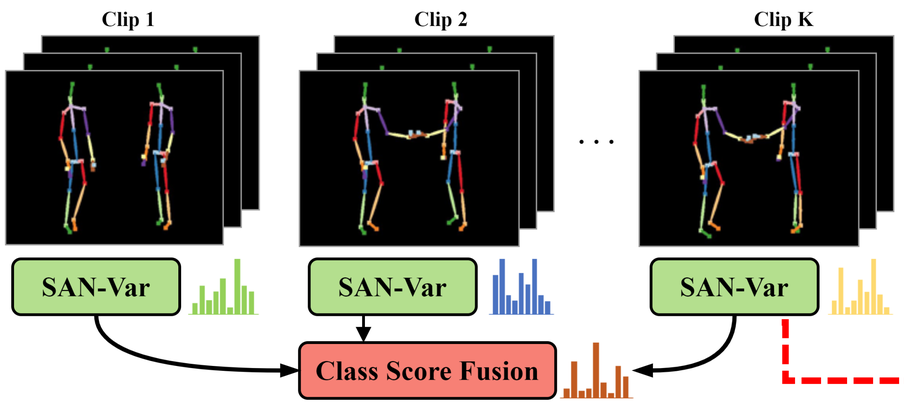
$\mathbf{x}_p \in \mathbb{R}^{F \times J' \times C}$

$\mathbf{x}^{(s)}_{(p,ff)} \in \mathbb{R}^{F \times J \times C'}$

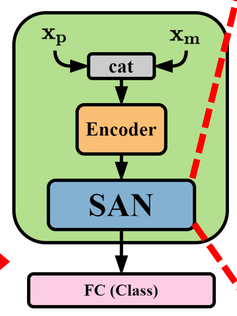
CNN: Extracts local spatio-temporal information



Self-Attention Network Variants

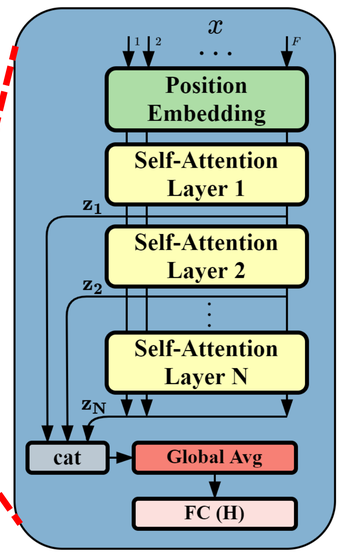


Overall network
- Temporal segments are fed to each SAN-Variant



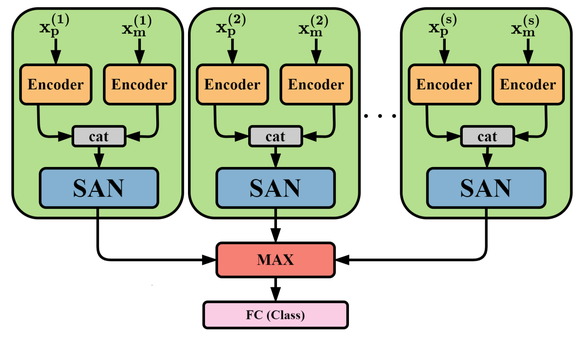
SAN-V1
Baseline

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

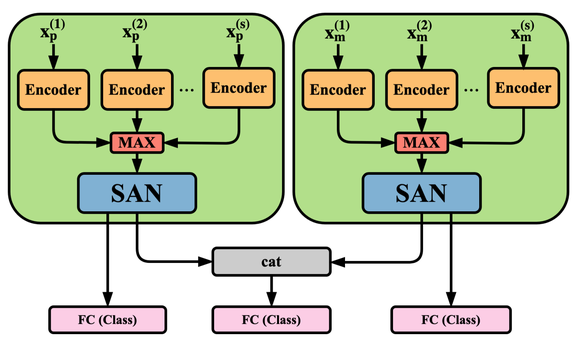


SAN block

- Extract all pairwise correlations (attentions) among frames
- Long & short range correlations



SAN-V2
SA for diff. subjects



SAN-V3
SA for diff. modalities

Experimental Results

NTU			Kinetics-400		
Evaluation	CS	CV	Evaluation	Top-1	Top-5
H-RNN	59.1	64.0	Feature Enc.	14.9	25.8
Ensemble TS-LSTM	74.6	81.3	Deep LSTM	16.4	35.3
VA-LSTM	79.4	87.6	Temporal Conv.	20.3	40.0
ST-GCN	81.5	88.3	ST-GCN	30.7	52.8
HCN	86.5	91.9	-	-	-
SAN V2 +FF	80.3	85.2	SAN V2 +FF	29.8	47.5
SAN V2 + CNN	85.9	91.7	SAN V2 + CNN	33.3	52.1
TS-SAN V2 (L2H2)	86.7	92.1	TS-SAN V2 (L2H2)	34.4	54.1
TS-SAN V3 (L4H8)	86.8	92.4	TS-SAN V3 (L4H8)	34.5	54.5
TS-SAN V2 (L4H8)	**87.2**	**92.7**	TS-SAN V2 (L4H8)	**35.1**	**55.7**

Visualization