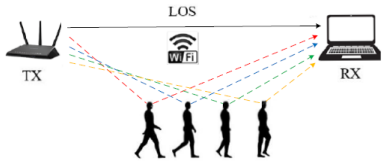
The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI21)

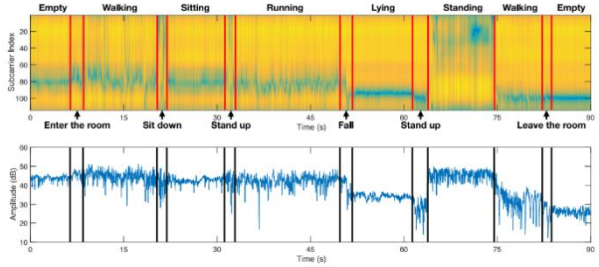
Bing Li[1], Wei Cui[2], Wei Wang[1], Le Zhang[2], Zhenghua Chen[2], Min Wu[2]
[1] University of New South Wales, Australia
[2] Institute for Infocomm Research, Agency for Science, Technology and Research(A*STAR), Singapore.

Two-Stream Convolution Augmented Transformer for Human Activity Recognition

Motivation

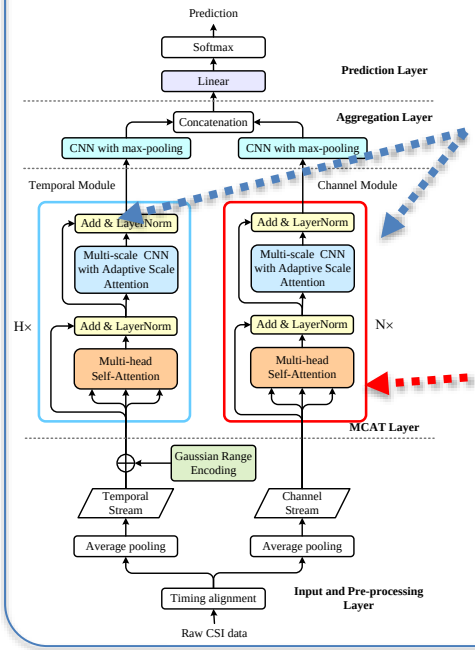
- Recognizing human actions using WiFi
  - Human actions incur multi-path and the fading effect on WiFi signals





  - Applications: healthcare and IoT systems, surveillance system

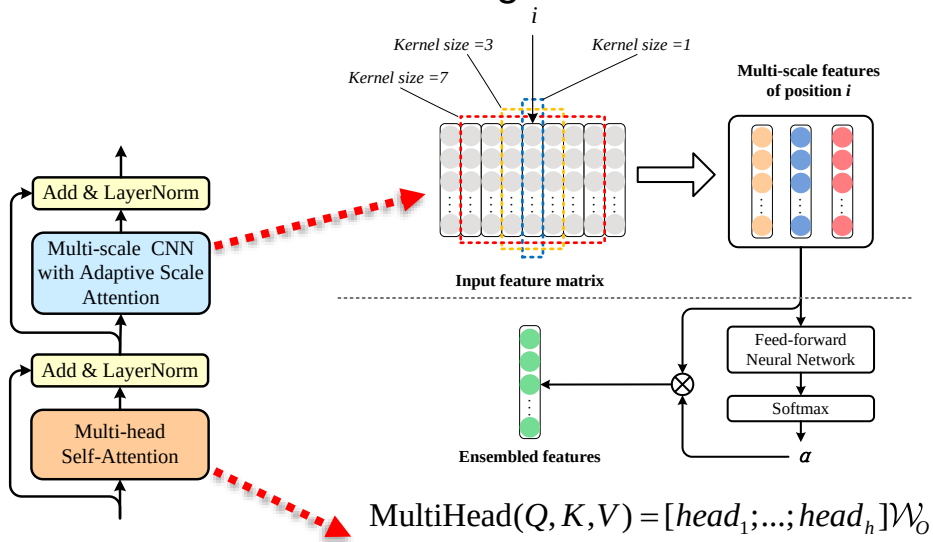
Model



1. Two-tower structure, each tower is in charge of the channel stream or the temporal stream to extract both time-over-channel and channel-over-time features

2. The MCAT model adopts a multi-scale convolution block to capture range-based patterns.

Multi-scale Convolution Augmented Transformer



$$\text{MultiHead}(Q,K,V) = [head_1;...;head_h]\mathcal{W}_O,$$

$$head_i = \text{Attention}(X\mathcal{W}_Q^i, X\mathcal{W}_K^i, X\mathcal{W}_V^i)$$

Experiment

The recognition accuracy (%) comparison on evaluation datasets

Datasets	S-RF	S-HMM	LSTM	CNN	ABLSTM	THAT
Office Room	75.3	79.7	91.4	96.4	97.1	**98.2**
Activity Room	80.0	75.0	89.7	94.3	95.6	**98.4**
Meeting Room	84.7	83.4	90.6	96.2	96.8	**99.0**
Activity+Meeting	82.6	80.5	90.1	95.4	95.9	**98.6**

Ablation study results compared with the full THAT model

Model	Office Room		Activity Room		Meeting Room		Activity+Meeting	
	Accuracy (%)	Δ	Accuracy (%)	Δ	Accuracy (%)	Δ	Accuracy (%)	Δ
THAT	**98.2**	-	**98.4**	-	**99.0**	-	**98.6**	-
- Gaussian Range Encoding	97.9	-0.4	97.9	-0.5	98.5	-0.5	98.2	-0.4
- Gaussian Range Encoding (+ PE (Vaswani et al. 2017))	91.1	-7.2	84.3	-14.1	90.3	-8.7	87.8	-10.8
- Multi-scale CNN	95.3	-3.0	97.4	-1.0	97.0	-2.0	97.3	-1.3
- Multi-scale CNN (+ PFFN)	97.2	-1.1	97.2	-1.2	98.1	-0.9	97.7	-0.9
- Temporal Module	92.0	-6.3	95.2	-3.2	97.5	-1.5	95.9	-2.7
- Channel Module	93.8	-4.5	97.7	-0.7	98.3	-0.7	98.0	-0.6