Poster: Multimodal ConvTransformer For Human Activity Recognition

Syed Tousiful Haque
Computer Science Department
Texas State University
San Marcos, Texas, USA
bgu9@txstate.edu

Anne H. H. Ngu
Computer Science Department
Texas State University
San Marcos, Texas, USA
angu@txstate.edu

***Abstract*—Recognition of human activity is crucial in emergency healthcare applications. Multi-modal deep learning learning algorithms are gaining attention for Human Activity Recognition (HAR) due to their success in various domains. The application of multi-modal learning to HAR continues to present challenges, particularly in addressing noisy data and achieving effective fusion of information from disparate modalities. We propose a new Multimodal-ConvTransformer (CT-HAR) that aims to efficiently extract and fuse complementary spatial and temporal information. Experiments conducted on the public UTD-Mhad and Berkley-Mhad datasets demonstrate significant performance enhancements, with CT-HAR achieving accuracy rates of 89.81% and 85.69% on these datasets, respectively.**

***Index Terms*—Human Activity Recognition, Multimodal Deep learning, Transformer**

I. Introduction

Human activity recognition(HAR) is used to detect and classify human activities under appropriate labels. Human activities are complex and change temporally. In recent times, due to the pervasive use of wearable devices by people, human activity data has become a common thing. Inertial sensors placed in a wearable device can provide linear and angular motion information of a particular joint of a human body. Although human activity data is available in large quantities, the common data modality provided is mainly of temporal information (time series) with little to no spatial information. Thus, HAR with only inertial data doesn't show high accuracy. Skeleton data captured from a Kinect camera can provide spatial and temporal information to support activity recognition. Deep learning models have increasingly been used in the field of HAR. Convolutional neural network (CNN), recurrent neural network (RNN), long-short-term memory (LSTM), and recently Transformer and Convolutional-Transformer (ConvTransformer) have been used for activity recognition. Despite Transformer and ConvTransformer having superior performance in HAR, combining information from multiple modalities remains a challenge. Some fusion techniques like TokenFusion [7], Cross-view Fusion [2], Mid-Fusion [3] were proposed to fuse multimodal information using a Transformer. However, these multimodal fusion methods are not effective in fusing complementary human activity information from different modalities as shown in their models with a large number of parameters with sub-par accuracy. To solve this problem, we propose a Multimodal ConvTransformer (CT-HAR) that uses Temporal Feature Fusion to combine information from both inertial and skeleton modalities.

The basic idea of Multimodal ConvTransformer is to use Convolution Neural Network and Transformer together to extract temporal and spatial information. A combination of Convolutional block and Transformer block extracts spatial and temporal information from skeleton data and a separate Transformer block extracts temporal information from inertial data. The temporal information from both modalities are then fused using Temporal Feature Fusion.

To show the advantage of the proposed method, we conducted experiments on two HAR datasets named UTD-Mhad and Berkley-Mhad. CT-HAR obtained competitive performance in these experiments and showed great efficiency. Especially, it improves the accuracy by 14.44%, 5.63%, and 10.92% over CrossVit [2], MidFusion [3] and TokenFusion [7] on UTD-Mhad dataset and by 8.12%, 6.8%, and 10.32% on Berkley-Mhad.

II. Overview of CT-HAR architecture

The CT-HAR is made up of three important parts that help it to analyze both spatial and temporal information effectively. The first part is the spatial module, represented by the Spatial Block in Fig 1. This module is in charge of dealing with the spatial details found in skeleton data. It uses two 2-dimensional convolution layers that are good at understanding the relationships between nearby joints. These layers have a special property called translation invariance inductive bias, making them particularly effective at processing spatial information. The output from this spatial module then moves on to the temporal block, which contains two Transformer encoder modules.

Fig. 1 shows structure of the encoder module within the temporal block. This module has the same architecture of Transformer encoder proposed in [5]. There are two separate temporal modules, one for skeleton data and one for inertial data. Temporal feature fusion involves the concatenation of outputs from corresponding encoders linked to skeleton data and inertial data. This concatenated output serves as the input for the subsequent encoder associated with skeleton data. This integration of spatial and temporal features is crucial

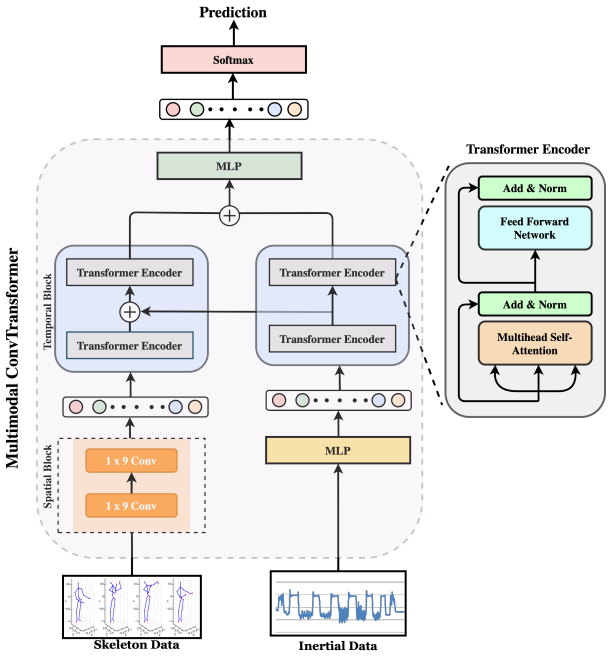

Fig. 1. Multimodal ConvTransformer (CT-HAR) Architecture

TABLE I
PERFORMANCE COMPARISON ON THE UTD-MHAD DATASET. S: SKELETON, D: DEPTH, I: INERTIAL

Method	Modality Combination	Accuracy(%)
JTM [6]	S	85.81
TokenFusion [7]	I + S	78.89
CrossViT [2]	I + S	75.37
MidFusion [3]	I + S	84.18
CT-HAR	I + S	**89.81**

for the Transformer's ability to understand and analyze the relationship between data of different modalities.

We used two public HAR datasets: UTD-Mhad [1] and Berkley-Mhad [4]. UTD-Mhad had 861 samples from 27 classes of different activities and Berkley-Mhad had 659 samples from 11 classes of different activities. We divided the dataset into the train/test split ratio as proposed in the dataset papers [1], [4] for both datasets. All experiments where conducted for 250 epoch with SGD optimizer and used a learning rate of 0.0025. The embedding dimension for the Transformer encoder layer was set to 32.

III. DISCUSSION

Table I and II show the experimental results. We compared the performance of CT-HAR with other Multimodal Transformer models, each employing a different fusion technique and also with Joint Trajectory Map based Convolutional Network(JTM) [6]. CT-HAR demonstrated superior performance compared to CrossVit [2], MidFusion [3], TokenFusion [7] and JTM [6] methodologies on the UTD-Mhad dataset, achieving an accuracy of 89.81%. This represents a consecutive gain of 14.44%, 5.63%,10.92% and 4.00% accuracy over the aforementioned methods, respectively. The increased accuracy of CT-HAR is primarily due to the Temporal Feature Fusion.

TABLE II
PERFORMANCE COMPARISON ON THE BERKLEY-MHAD DATASET. S: SKELETON, D: DEPTH, I: INERTIAL

Method	Modality Combination	Accuracy(%)
TokenFusion [7]	I + S	78.89
CrossVit [8]	I + S	75.37
Midfusion [3]	I + S	77.57
CT-HAR	I + S	**85.69**

Similarly, on Berkley-Mhad the accuracy for CT-HAR improved by 8.12%, 6.8%, and 10.32% compared to MidFusion [3], TokenFusion [7] and CrossVit [2] as it achieves a 85.69% accuracy. These results highlight that CT-HAR can generalize over different datasets.

IV. CONCLUSION

The CT-HAR framework adeptly extracts and synthesizes information from diverse modalities, demonstrating notable enhancements in accuracy across two multimodal HAR datasets. The integration of Convolutional Layers and Transformer encoders facilitates the effective extraction of spatial and temporal information and Temporal Feature Fusion is instrumental in generating representations that are rich in both spatial and temporal information.

ACKNOWLEDGMENT

We thank the National Science Foundation for funding the research under the Smart and Connected Health Program (NSF-SCH-21223749) at Texas State University.

REFERENCES

[1] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 ICIP*, pages 168–172. IEEE, 2015.

[2] Momal Ijaz, Renato Diaz, and Chen Chen. Multimodal transformer for nursing activity recognition. In *Proceedings of the IEEE/CVF CVPR*, pages 2065–2074, 2022.

[3] Jingcheng Li, Lina Yao, Binghao Li, and Claude Sammut. Distilled mid-fusion transformer networks for multi-modal human activity recognition. *arXiv preprint arXiv:2305.03810*, 2023.

[4] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *2013 IEEE (WACV)*, pages 53–60. IEEE, 2013.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 30, 2017.

[6] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. Action recognition based on joint trajectory maps using convolutional neural networks. In *Proceedings of the 24th ACM Multimedia*, pages 102–106, 2016.

[7] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF CVPR*, pages 12186–12195, 2022.

[8] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, pages 3333–3343, 2022.