Language Knowledge-Assisted Representation Learning for HAR

Motivation

- Existing studies **incorporate prior knowledge** to aid potential representation learning for better performance
- LA-GCN propose GCN using large-scale language model (LLM) assistance
  - LLM is mapped into priori global relationship (GPR) & priori categorical relationship (CPR)
    - GPR : **new “bone” representation**, aim to emphasize essential node information
    - CPR: simulate **category prior knowledge in human brain region**
  - Propose multi-hop attention GCN: improving the efficiency of message passing between nodes

Introduction

- Motivation
  - Temporoparietal is stimulated to activate corresponding brain area associated with action
  - Prediction is accomplished by reasoning based on a priori knowledge and goals
- GPR & CPR
  - GPR: guide **new skeleton** to model critical information from data level
  - CPR: compose **priori consistency-assisted classification (PC-AC)** to help model learning based on feature with enhanced semantic relationship

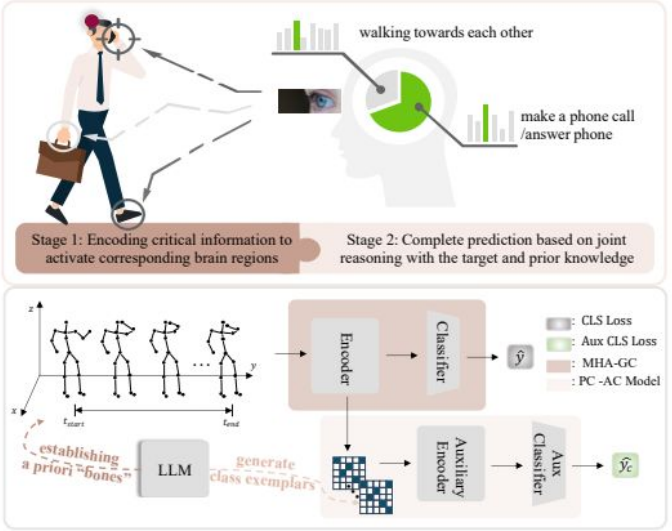


Fig. 1: Schematic of LA-GCN concept. The top half of this figure shows two brain activity processes when humans perform action recognition. The bottom half shows the proposed multi-task learning process. The knowledge of the language model is divided into global information and category information to simulate the a priori knowledge used in human reasoning to aid the model. The encoder infers the correlation between joints and thus refines the topology using contextual information.

Global Prior Relation (GPR)

- Step1: Extract text feature for each class of action label and all joints using BERT
  - In BERT
    - Input: **a sentence**
    - Output (after Pooler): semantic of sentence (can use for text classification)
  - In this work
    - Input: combination **class name + joint name** (e.g. [joint] function in [class])
    - Output: feature **J** ∈ $R^{ClsxVxC}$
    - With **Cls** actions, **V** nodes, C=256 feature dim

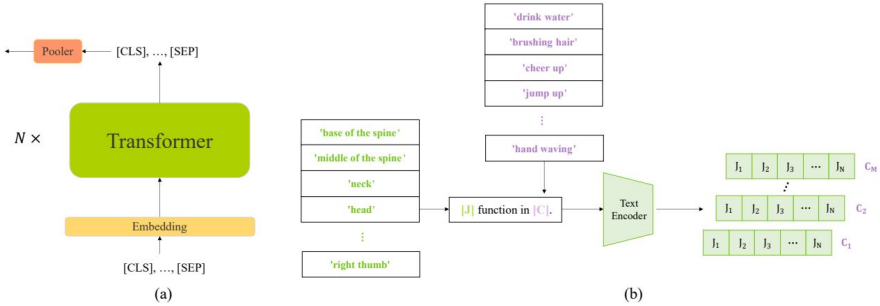


Fig. 2: Extraction of text features. Subfigure (a) is Bert's architecture. (b) Our method uses the learned text encoder to extract text features by embedding the names of classes [C] and the names of all joints [J] of the target dataset.

Global Prior Relation (GPR) (2)

- Step2: obtain GPR topology graph containing the **semantic knowledge** guide skeleton representation generation
  - Find the centroids (CoCLS) of **each node**
    - $\mathbf{J} \in R^{ClsxVxC}$ average class => $\mathbf{J}^{CoCLS} \in R^{NxC}$
  - Calculated correlation
    - $\mathbf{J}^{CoCLS}$ x $\mathbf{J}^{CoCLS}$ -> GPR Graph $\in R^{VxV}$

  => GPR Graph obtained

- **CPR**
  - Calculated correlation
    - **J** x **J** -> GPR Graph $\in R^{VxVxCls}$

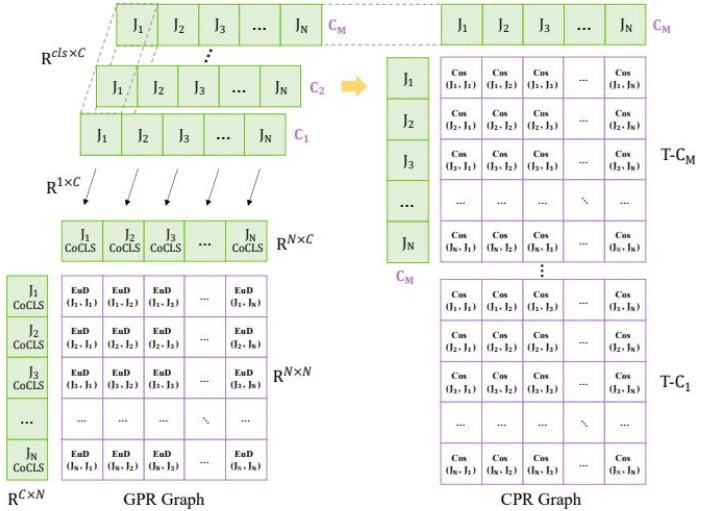


Fig. 3: Summarize our approach to generate prior topologies. GPR Graph is obtained by computing the class centers of the joints and computing correlations between the node feature of each action to obtain the CPR Graph.

Priori Skeleton Representation

Objective: new skeleton representation using GPR Graph

- Extract all “bone” data of NTU 60, bone have standard deviation
- Define the “bone” selection function as g(·)
  - Ex: $B\sim = g_{min}(B_{std})$
-

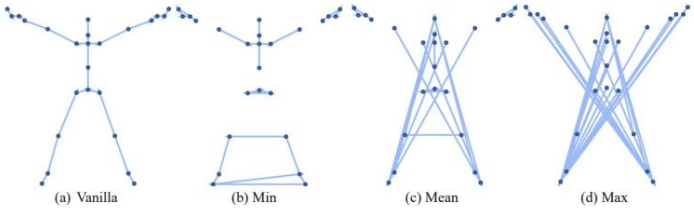
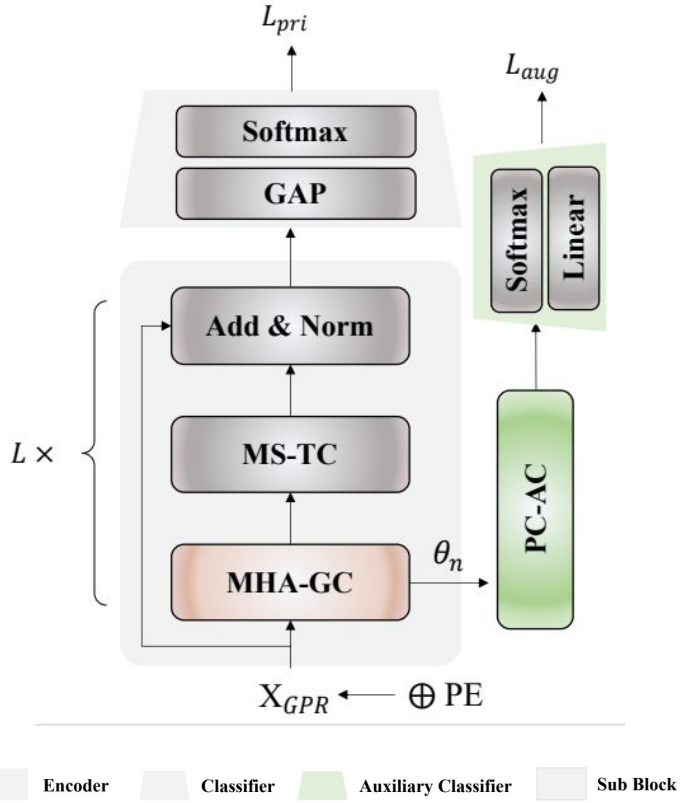


Fig. 4: Multimodal representation of the skeleton. The arrows depict the “bone” vector from the source to the target joint. (a) is the bone matrix [16]. (b), (c) and (d) is our minimum, mean, and maximum std summation matrix, respectively. It was found that nearly half of the skeletons in (b) are consistent with (a) but contain more detailed relationships, and (c) implicitly contains information about the connections between body parts, such as hands, torso and hands, torso and legs, and legs. Where (b) is used as our new “bone” representation.

LA-GCN architecture: Input data

- Architecture
  - Input data
  - Encoder block
  - SubBlock
  - Classifier
- Input data
  - Skeleton sequence X with
    - Weighted using GPR Graph
    - Position embedding (abstraction position)

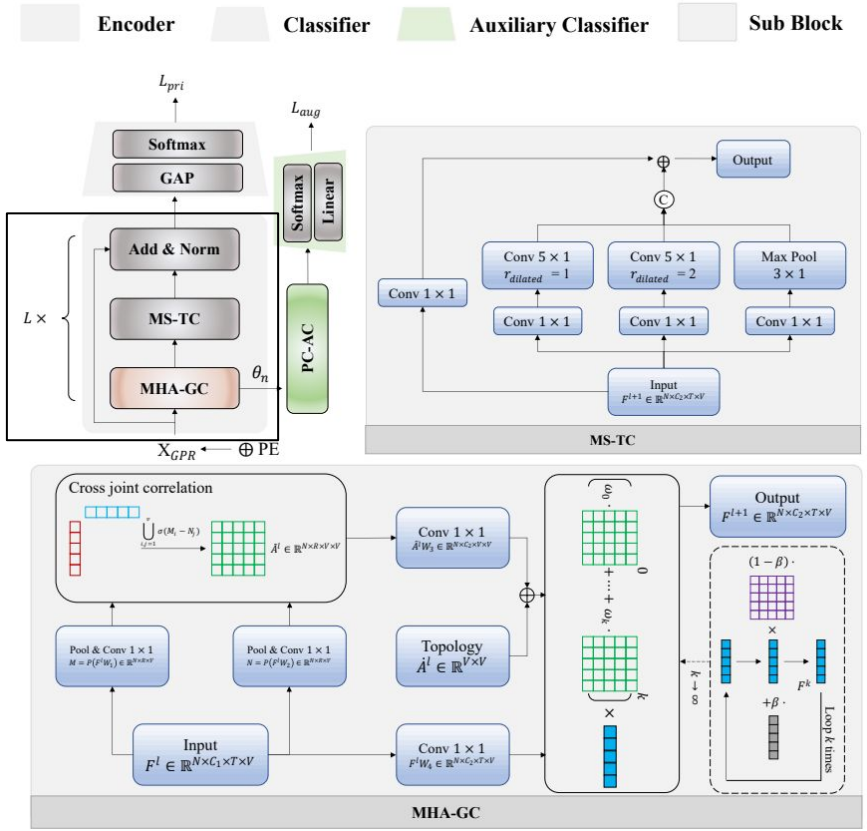
$$F^{(0)} = \mathrm{X}_{GPR} W_0 + PE,$$



LA-GCN architecture: Encoder Block

- MS-TC:
  - Multiscale Temporal encoder with different time lengths
- MHA-GC
  - Multi-hop attention GCN for **inter-node relationship modeling**
    1. First-order neighborhood $A\tilde{}^{l}$
    2. Shared topology $A^{\cdot l} \in R^{V \times V}$
    3. Attention $\bar{A}^l = \dot{A}^l + \gamma \tilde{A}^l W_3,$
    4. Feature aggregation function

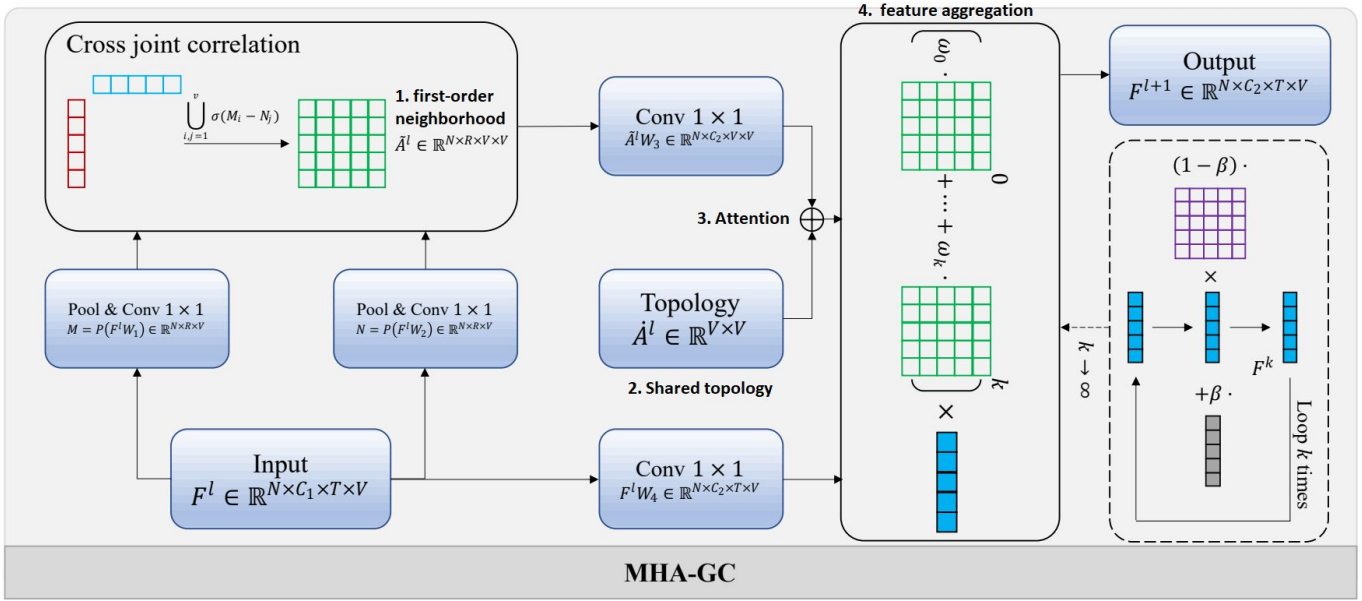
$$F^{l+1} = \sigma(\bar{A}^l F^l W_4^l),$$



LA-GCN architecture: Encoder Block (2)

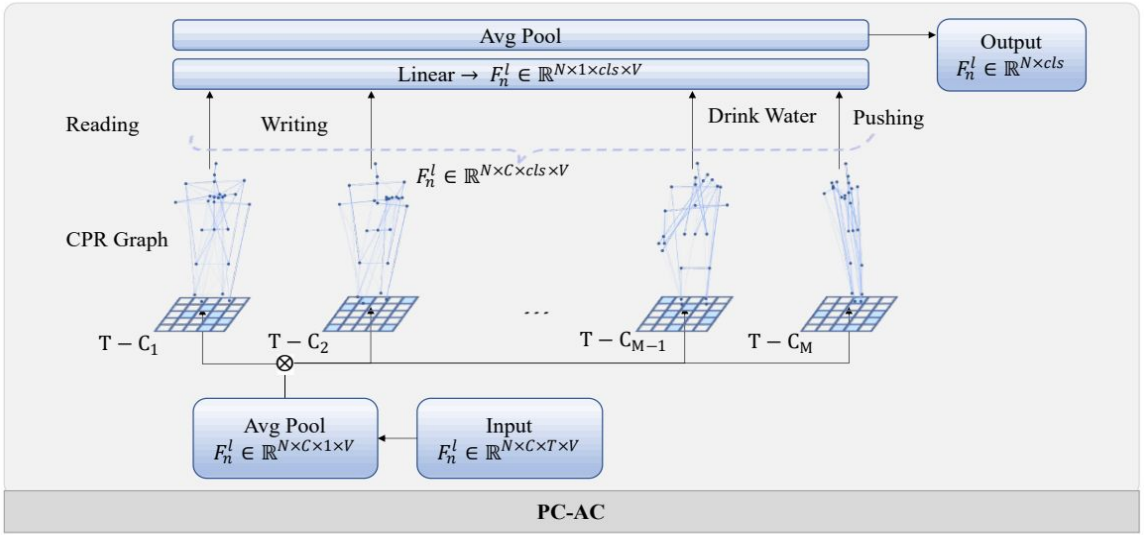
MHA-GC: Multi-hop attention GCN for **inter-node relationship modeling**

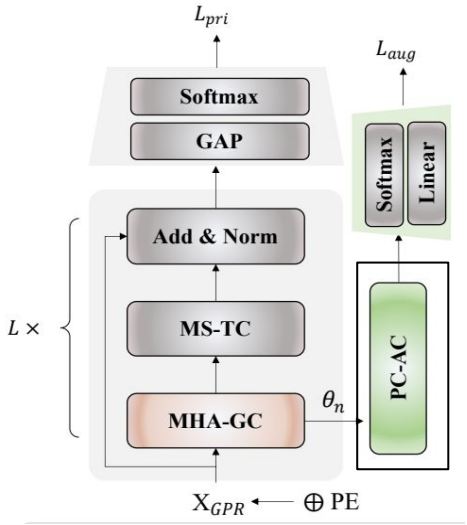
1. First-order neighborhood $A\tilde{}^{l}$
2. Shared topology $A^{\cdot l} \in R^{V \times V}$
3. Attention $\bar{A}^l = \dot{A}^l + \gamma \tilde{A}^l W_3,$
4. Feature aggregation function $F^{l+1} = \sigma(\bar{A}^l F^l W_4^l),$



LA-GCN architecture: SubBlock

- Objective: design a class relationship topology graph containing a priori node relationships to perform feature aggregation
  - Given input: F ∈ $R^{NxCxTxV}$ => Avg pool: F ∈ $R^{NxCx1xV}$
  - Combine with CPR: ∈ $R^{VxVxCls}$
  - Output: F ∈ $R^{NxCxClsxV}$ => Linear + Avg => Output: F ∈ R^{NxCls}





Result

Dataset	Model	Metric Name	Metric Value	Global Rank
NTU RGB+D	LA-GCN	Accuracy (CV)	97.2	# 5
		Accuracy (CS)	93.5	# 2
		Ensembled Modalities	6 (non-standard, w/o motion modalities)	# 12
NTU RGB+D 120	LA-GCN	Accuracy (Cross-Subject)	90.7	# 1
		Accuracy (Cross-Setup)	91.8	# 1
		Ensembled Modalities	6 (non-standard, w/o motion modalities)	# 14
N-UCLA	LA-GCN	Accuracy	97.6	# 1

TABLE 1: The comparison of Top-1 accuracy (%) on the NTU RGB+D [31] benchmark.

Method	X-Sub	X-View
VA-LSTM [25]	79.4	87.6
ST-GCN [16]	81.5	88.3
AS-GCN [51]	86.8	94.2
2s-AGCN [13]	88.5	95.1
AGC-LSTM [39]	89.2	95.0
Directed-GNN [52]	89.9	96.1
ST-TR [53]	90.3	96.3
Shift-GCN [40]	90.7	96.5
DC-GCN+ADG [54]	90.8	96.6
Dynamic-GCN [12]	91.5	96.0
MS-G3D [11]	91.5	96.2
DDGCN [55]	91.1	97.1
MST-GCN [56]	91.5	96.6
EfficientGCN [7]	92.1	96.1
CTR-GCN [8]	92.4	96.8
Info-GCN [6]	93.0	97.1
LA-GCN (ours)	**93.5**	**97.2**

TABLE 3: The comparison of Top-1 accuracy (%) on the NW-UCLA [33] benchmark.

Methods	Top1
Lie Group [57]	74.2
Actionlet ensemble [58]	76.0
HBRNN-L [59]	78.5
Ensemble TS-LSTM [38]	89.2
AGC-LSTM [39]	93.3
4s Shift-GCN [40]	94.6
DC-GCN+ADG [54]	95.3
CTR-GCN [8]	96.5
InfoGCN [6]	97.0
LA-GCN (ours)	**97.6**

TABLE 2: The comparison of Top-1 accuracy (%) on the NTU RGB+D 120 [32] benchmark.

Method	X-Sub	X-Set
Part-Aware LSTM [31]	26.3	25.5
ST-LSTM [37]	55.7	57.9
RotClips+MTCNN [36]	62.2	61.8
ST-GCN [16]	70.7	73.2
2s-AGCN [13]	82.9	84.9
SGN [50]	82.9	84.9
ST-TR [53]	85.1	87.1
Shift-GCN [40]	85.9	87.6
MS-G3D [11]	86.9	88.4
Dynamic-GCN [12]	87.3	88.6
MST-GCN [56]	87.5	88.8
EfficientGCN [7]	88.7	88.9
CTR-GCN [8]	88.9	90.6
Info-GCN [6]	89.4	90.7
LA-GCN (Joint)	86.5	88.0
LA-GCN (Joint+Bone)	89.7	90.9
LA-GCN (4 ensemble)	89.9	91.3
LA-GCN (6 ensemble)	**90.7**	**91.8**

Ablation Study: Effectiveness of PC-AC

TABLE 4: (i) The comparison of accuracy without and with PC-AC, where the λ of L_{aug} is 0.2. (ii) Contribution of different text prompts.

Methods	Top1
w/o L_{aug}	84.9
L_{total}	**86.1**$^{\uparrow 1.2}$
p1: [J] function in [C].	85.6$^{\uparrow 0.7}$
p2: What happens to [J] when a person is [C]?	85.8$^{\uparrow 0.9}$
p3: What will [J] act like when [C]?	**86.1**$^{\uparrow 1.2}$
p4: When [C][J] of human body.	85.5$^{\uparrow 0.6}$
p5: When [C] what will [J] act like?	85.7$^{\uparrow 0.8}$
p6: When a person is [C], [J] is in motion.	85.5$^{\uparrow 0.6}$

Ablation Study: Improved Interaction Between Neighbor Nodes

TABLE 5: (i) The comparison of accuracy without and with PC-AC, where the λ of L_{aug} is 0.2. (ii) Contribution of different text prompts.

Methods	Top1
A	86.1
$\bar{A}$	**86.5**
1-hop : A	86.1
2-hop	86.3
3-hop	OOM
2-hop^{1st}	86.1
3-hop^{1st}	86.2
4-hop^{1st}	**86.5**
5-hop^{1st}	85.9