3Mformer: Multi-order Multi-mode Transformer for Skeletal Action Recognition

Lei Wang[1,2] Piotr Koniusz[2,1]

[1]Australian National University
[2]Data61/CSIRO

May 28, 2023

3Mformer: Multi-order Multi-mode Transformer for Skeletal Action Recognition

Lei.Wang@data61.csiro.au[1,2] Piotr.Koniusz@data61.csiro.au[2,1]

[1]Australian National University [2]Data61/CSIRO

Australian National University DATA 61 CSIRO

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

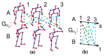
Motivation

Existing GCN-based action recognition models:

- represent human body joints based on physical connectivity
- limited receptive fields & one-/few-hop neighbourhood aggregation
- ignore dependency between body joints non-connected by body parts

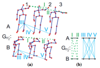
Human actions are associated with interaction groups of skeletal joints:

- the impact of groups of joints on each action differs
- the degree of influence of each joint should be learned
- design a better model for skeleton data (topology of skeleton graph)



Inspired by our tensor representations[a]:

- *sequence compatibility kernel* (SCK) & *dynamics compatibility kernel* (DCK)
- incorporate multi-modal inputs & compactly capture complex interplay
- operate on subsequences / capture local-global interplay of correlations

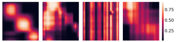


[a]Koniusz, P., Wang, L., & Cherian, A. (2021). **Tensor representations for action recognition.** *IEEE TPAMI*, 44(2), 648-665.

Key ideas

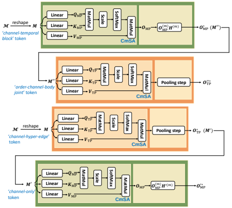
We use skeletal hypergraph, hypergraph captures higher-order relationships by hyper-edges.
Given $\mathcal{M}\in\mathbb{R}^{I_1\times I_2\times\ldots\times I_r}$, we perform mode-$m$ matricization to obtain $\mathbf{M}\equiv\mathcal{M}_{(m)}^{\top}\in\mathbb{R}^{(I_1\ldots I_{m-1}I_{m+1}\ldots I_r)\times I_m}$ to form coupled-token: 'channel-temporal block', 'channel-body joint', 'channel-hyper-edge (any order)', and 'channel-only' pairs.
Coupled-mode Self-Attention (CmSA):

- show diagonal / vertical patterns
- patterns are consistent with the pattens of attention matrices found in standard Transformer, *e.g.*, NLP



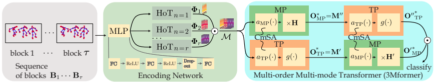
We propose a **Multi-order Multi-mode Transformer (3Mformer)**, which uses coupled-mode tokens to jointly learn various higher-order motion dynamics. Two basic building modules:

- Multi-order Pooling (MP)
  - combine information flow **block-wise**
  - **various coupled-mode** tokens help improve results
  - **different focus** of each attention mechanism
- Temporal block Pooling (TP)
  - each sequence may contains a different number of blocks
  - aggregates via popular pooling, *e.g.*, rank-, first-, second- or higher-order pooling



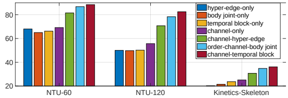
We form **multi-head** CmSA.

The pipeline: further details



- Each sequence is split into τ temporal blocks $\mathbf{B}_1, ..., \mathbf{B}_\tau$
- Each block is embedded by a simple MLP into $\mathbf{X}_1, ..., \mathbf{X}_\tau$
- $\mathbf{X}_1, ..., \mathbf{X}_\tau$ are passed to HoTs ($n=1, ..., r$) for feature tensors $\mathbf{\Phi}_1, ..., \mathbf{\Phi}_\tau$
- Subsequently concatenated by $\odot$ along the hyper-edge mode into tensor $\mathbf{M}$
- **3Mformer contains two complementary branches**: MP→TP & TP→MP
- Outputs are concatenated by $\odot$ and passed to the classifier
- MP & TP perform attention with the so-called **coupled-mode tokens**
- MP contains **weighted pooling along hyper-edge mode** by learnable matrix $\mathbf{H}$ (& $\mathbf{H}'$ in another branch).
- TP contains **block-temporal pooling** denoted by $g(\cdot)$ to capture block-temporal order with pooling

Results



	Method	Venue	NTU-60		NTU-120		Kinetics-Skeleton	
			X-Sub	X-View	X-Sub	X-Set	Top-1	Top-5
Graph-based	ST-GCN	AAAI'18	81.5	88.3	70.7	73.2	30.7	52.8
	AS-GCN	CVPR'19	86.8	94.2	78.3	79.8	34.8	56.5
	2S-AGCN	CVPR'19	88.5	95.1	82.5	84.2	36.1	58.7
	NAS-GCN	AAAI'20	89.4	95.7	-	-	37.1	60.1
	Sym-GNN	TPAMI'22	90.1	96.4	-	-	37.2	58.1
	Shift-GCN	CVPR'20	90.7	96.5	85.9	87.6	-	-
	MS-G3D	CVPR'20	91.5	96.2	86.9	88.4	38.0	60.9
	CTR-GCN	ICCV'21	92.4	96.8	88.9	90.6	-	-
	InfoGCN	CVPR'22	93.0	97.1	89.8	91.2	-	-
	PoseConv3D	CVPR'22	94.1	97.1	86.9	90.3	**47.7**	-
Hypergraph-based	Hyper-GNN	TIP'21	89.5	95.7	-	-	37.1	60.0
	SD-HGCN	ICONIP'21	90.9	96.7	87.0	88.2	37.4	60.5
Transformer-based	ST-TR	CVIU'21	90.3	96.3	85.1	87.1	38.0	60.5
	STST	ACM MM'21	91.9	96.8	-	-	38.3	61.2
	3Mformer (with max-pool, *ours*)		92.1	97.8	-	-	-	-
	3Mformer (with attn-pool, *ours*)		**94.2**	**98.5**	89.7	92.4	45.7	67.6
	3Mformer (with tri-pool, *ours*)		94.0	**98.5**	**91.2**	**92.7**	**47.7**	**71.9**
	3Mformer (with rank-pool, *ours*)		**94.8**	**98.7**	**92.0**	**93.8**	**48.3**	**72.3**

Motivation

- GCN-based
  - represent human body joints based on **physical connectivity**
  - **limited** receptive fields / one- or few-hop neighbourhood aggregation
  - ignore the dependency between body joints **non-connected** by body parts
- Human actions are associated with **interaction groups of skeletal joints**
  - the impact of groups of joints on each action differs
- Inspired by our tensor representations[1]:
  - *sequence compatibility kernel* (SCK) & *dynamics compatibility kernel* (DCK)
  - compactly **capture complex interplay**
  - operate on **subsequences** / capture the local-global interplay of correlations
  - incorporate **multi-modal inputs**

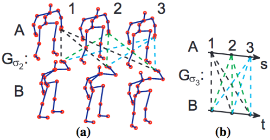


Figure 1: SCK

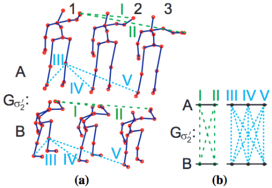


Figure 2: DCK

[1] Koniusz, P., Wang, L., & Cherian, A. (2021). **Tensor representations for action recognition**. *IEEE TPAMI*, 44(2), 648-665.

Motivation (cont.)

We propose to:

- use skeletal hypergraph
- Hypergraph captures higher-order relationships by hyper-edges
- Hyper-edges connect more than two nodes (body joints)

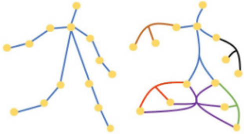


Figure 3: Skeletal graph & hypergraph.

Compared to GCN:

- encodes **first-**/**second-**/ **higher-**order hyper-edges
- set of body joints (**nodes**)/ **edges** between pairs of nodes/**hyper-edges** between triplets of nodes

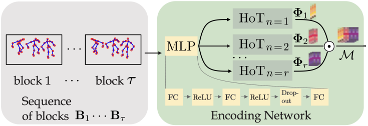


Figure 4: MLP+HoT branches

Concatenating HoT outputs of orders 1 to r across τ^2 blocks is *sub-optimal*.

- #hyper-edges of J joints **grows rapidly with order** r, *i.e.*, $\binom{J}{i}$ for $i = 1, ..., r$
- embeddings of the **highest order hyper-edges dominate lower orders**
- **long-range temporal dependencies** of features are insufficiently explored

[2] For brevity, we write that we have τ temporal blocks per sequence. In fact, τ varies.

Multi-order Multi-mode Transformer (3Mformer)

Given $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_r}$, we perform mode-$m$ matricization to obtain $\mathbf{M} \equiv \boldsymbol{\mathcal{M}}_{(m)}^{\top} \in \mathbb{R}^{(I_1 \ldots I_{m-1} I_{m+1} \ldots I_r) \times I_m}$ to form coupled-token.

- **Coupled-mode tokens**:
  - 'channel-temporal block' (Attention matrix $\mathbf{A}_{\mathrm{MP}} \in \mathbb{R}^{d'\tau \times d'\tau}$)
  - 'channel-body joint' ($\mathbf{A}_{\mathrm{TP}} \in \mathbb{R}^{rd'J \times rd'J}$)
  - 'channel-hyper-edge (any order)' ($\mathbf{A}_{\mathrm{TP}} \in \mathbb{R}^{d'N \times d'N}$ & $N = \sum_{m=1}^{r} \binom{J}{m}$)
  - and 'channel-only' ($\mathbf{A}_{\mathrm{MP}} \in \mathbb{R}^{d' \times d'}$) pairs
- **Coupled-mode Self-Attention (CmSA)**:
  - show diagonal / vertical patterns
  - patterns are consistent with the pattens of attention matrices found in standard Transformer, *e.g.*, NLP

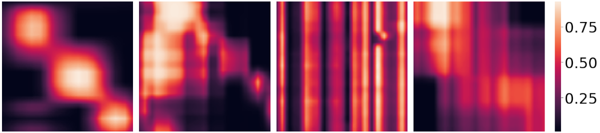


Figure 5: Visualization of attention matrices: 'channel-only', 'channel-hyper-edge', 'order-channel-body joint' & 'channel-temporal block' tokens.

Visualization of 3Mformer

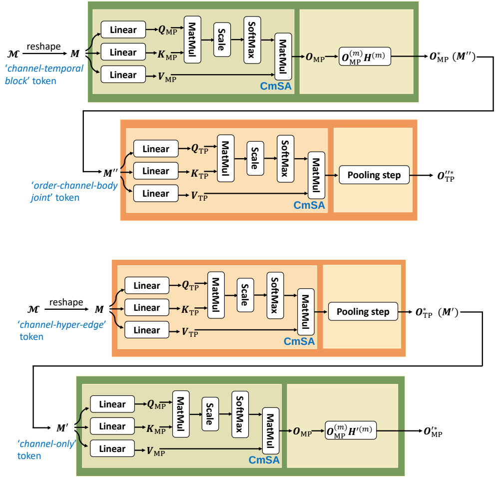


Figure 6: 3Mformer is a two-branch model: (a) MP→TP & (b) TP→MP.

Two basic building modules:

- Multi-order Pooling (MP)
  - combine information flow **block-wise**
  - **various coupled-mode** tokens help improve results
  - **different focus** of each attention mechanism
- Temporal block Pooling (TP)
  - each sequence may contains a different number of blocks
  - aggregates via popular pooling, *e.g.*, rank-, first-, second- or higher-order pooling

We also form our **multi-head** CmSA as in standard Transformer.

Pipeline: further details

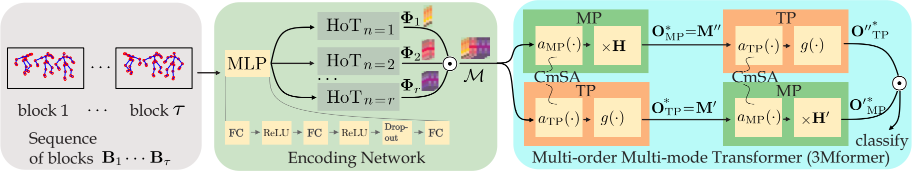


Figure 7: Pipeline overview.

- each sequence is split into τ temporal blocks $\mathbf{B}_1, ..., \mathbf{B}_\tau$
- each block is embedded by a simple MLP into $\mathbf{X}_1, ..., \mathbf{X}_\tau$
- $\mathbf{X}_1, ..., \mathbf{X}_\tau$ are passed to HoTs $(n=1, ..., r)$ for feature tensors $\mathbf{\Phi}_1, ..., \mathbf{\Phi}_\tau$
- subsequently concatenated by $\odot$ along the hyper-edge mode into tensor $\mathbf{M}$
- **3Mformer contains two complementary branches**: MP→TP & TP→MP
- outputs are concatenated by $\odot$ and passed to the classifier
- MP & TP perform attention with the so-called **coupled-mode tokens**
- MP contains **weighted pooling along hyper-edge mode** by learnable matrix $\mathbf{H}$ (and $\mathbf{H}'$ in another branch).
- TP contains **block-temporal pooling** denoted by $g(\cdot)$ to capture block-temporal order with pooling

Results & Discussions

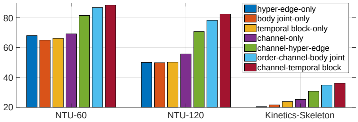


Figure 8: Single-mode *vs.* coupled-mode.

Table 1: NTU-60, NTU-120 & Kinetics-Skeleton.

	Method	Venue	NTU-60		NTU-120		Kinetics-Skeleton	
			X-Sub	X-View	X-Sub	X-Set	Top-1	Top-5
Graph-based	TCN	CVPRW'17	-	-	-	-	20.3	40.0
	ST-GCN	AAAI'18	81.5	88.3	70.7	73.2	30.7	52.8
	AS-GCN	CVPR'19	86.8	94.2	78.3	79.8	34.8	56.5
	2S-AGCN	CVPR'19	88.5	95.1	82.5	84.2	36.1	58.7
	NAS-GCN	AAAI'20	89.4	95.7	-	-	37.1	60.1
	Sym-GNN	TPAMI'22	90.1	96.4	-	-	37.2	58.1
	Shift-GCN	CVPR'20	90.7	96.5	85.9	87.6	-	-
	MS-G3D	CVPR'20	91.5	96.2	86.9	88.4	38.0	60.9
	CTR-GCN	ICCV'21	92.4	96.8	88.9	90.6	-	-
	InfoGCN	CVPR'22	93.0	97.1	89.8	91.2	-	-
	PoseConv3D	CVPR'22	94.1	97.1	86.9	90.3	**47.7**	-
Hypergraph-based	Hyper-GNN	TIP'21	89.5	95.7	-	-	37.1	60.0
	DHGCN	CoRR'21	90.7	96.0	86.0	87.9	37.7	60.6
	Selective-HCN	ICMR'22	90.8	96.6	-	-	38.0	61.1
	SD-HGCN	ICONIP'21	90.9	96.7	87.0	88.2	37.4	60.5
Transformer-based	ST-TR	CVIU'21	90.3	96.3	85.1	87.1	38.0	60.5
	MTT	LSP'21	90.8	96.7	86.1	87.6	37.9	61.3
	4s-GSTN	Symmetry'22	91.3	96.6	86.4	88.7	-	-
	STST	ACM MM'21	91.9	96.8	-	-	38.3	61.2
	3Mformer (with avg-pool, *ours*)		92.0	97.3	88.0	90.1	43.1	65.2
	3Mformer (with max-pool, *ours*)		92.1	97.8	-	-	-	-
	3Mformer (with attn-pool, *ours*)		**94.2**	**98.5**	89.7	92.4	45.7	67.6
	3Mformer (with tri-pool, *ours*)		94.0	**98.5**	**91.2**	**92.7**	**47.7**	**71.9**
	3Mformer (with rank-pool, *ours*)		**94.8**	**98.7**	**92.0**	**93.8**	**48.3**	**72.3**

Discussions:

- Single-mode *vs.* coupled-mode
- graph-based *vs.ours*:
  - AS-GCN/2S-AGCN
    - pairwise relationship
    - second-order
  - *ours*
    - higher-order
    - groups of body joints
  - 2nd-order HoT alone *vs.* NAS-GCN/Sym-GNN
- hypergraph-based *vs.ours*:
  - 3rd-order HoT alone *vs.* Hyper-GNN/SD-HGCN/Selective-HCN

Thank you!