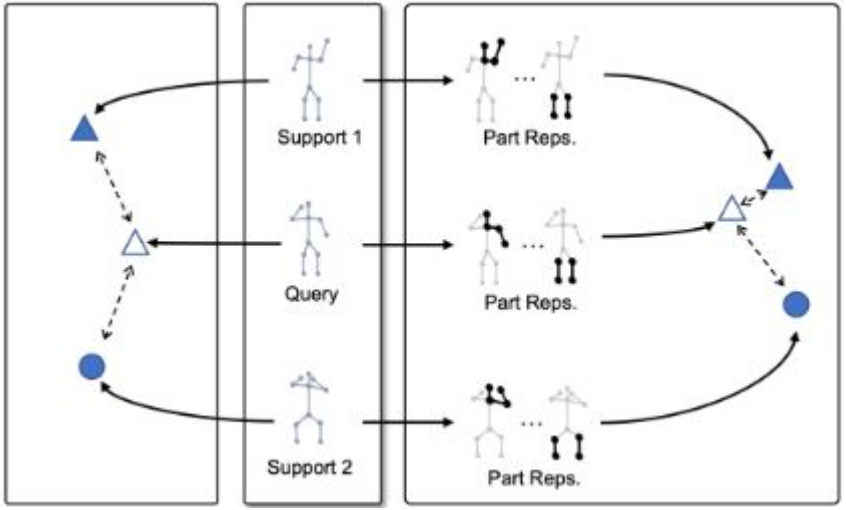
Part-aware Prototypical Graph Network for One-shot Skeleton-based Action Recognition

Sumario

1. Introdução 3
2. Objetivos 4
3. Metodologia ..... 5 - 10
4. Base de dados experimentadas 11
5. Resultados .....12 - 15
6. Conclusão 16
7. Dúvidas, sugestões e discussão 17

Introdução

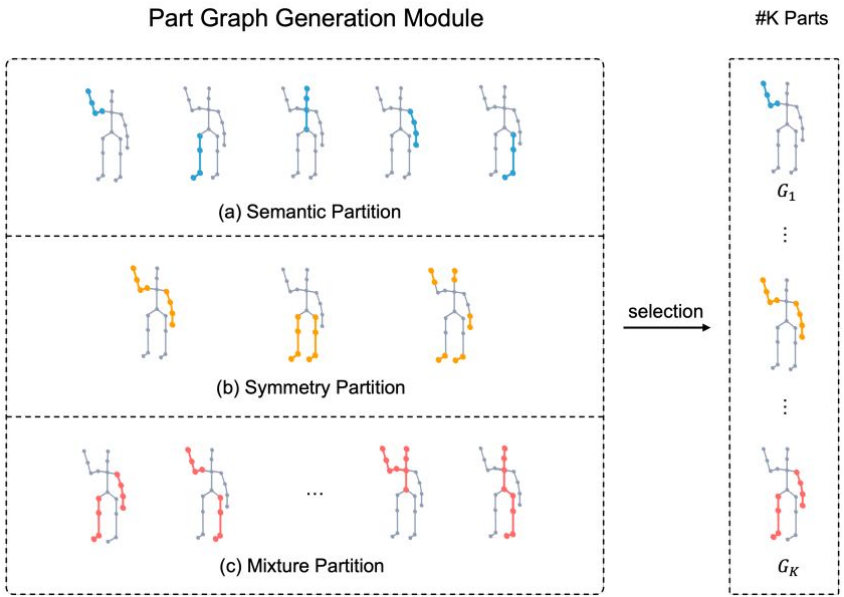
- Explorar espaço de juntas locais.
- *Metric-Learning*



Os métodos existentes (à esquerda) normalmente dependem da representação da pose completa para reconhecimento das ações. Por outro lado, o método proposto (à direita) adota um modelo de reconhecimento usando conjunto de juntas locais.

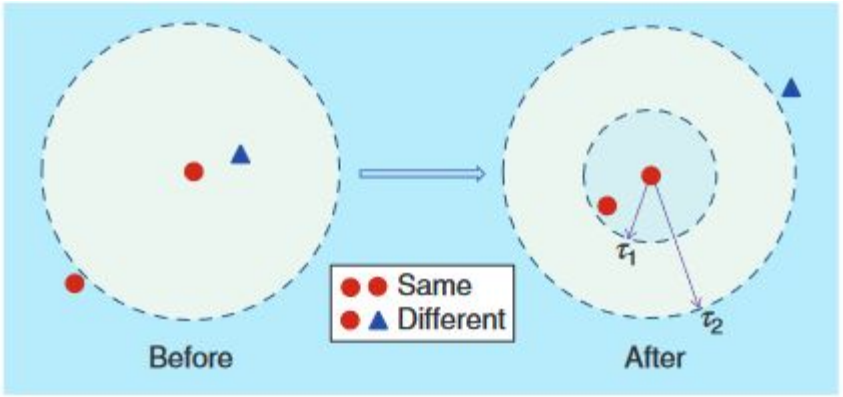
Objetivos

- Espaço de juntas locais.
  - Melhorias de reconhecimento usando *Metric Learning*.
  - Melhorias de reconhecimento usando *Graph Convolutional Networks* (GCN) em *few-shot learning*.
- Padrões de movimentação local traz melhorias



Exemplo de espaço de juntas locais.

Metodologia: *Deep Metric Learning*

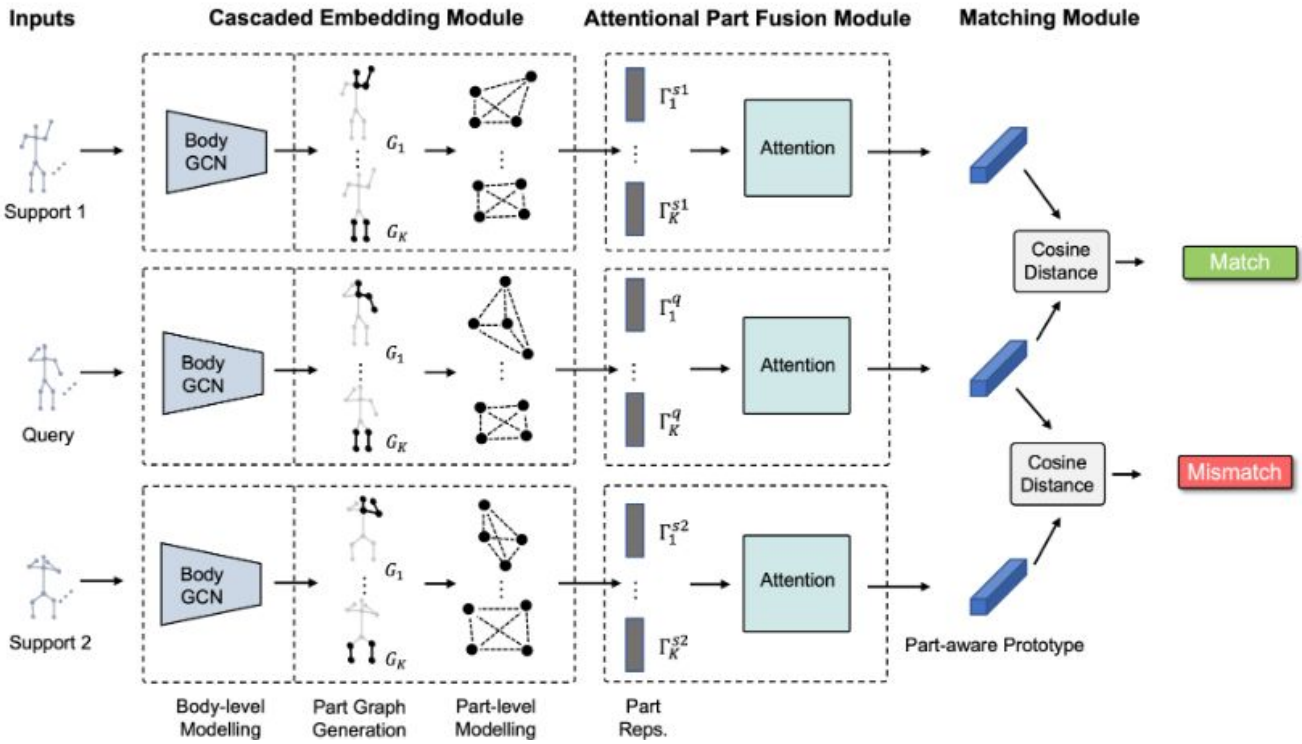


A ideia básica dos métodos *Deep Metric Learning* (DML) usando redes siamesas. Onde cada ponto no contexto do problema é uma ação.

Metodologia: *Metric learning*

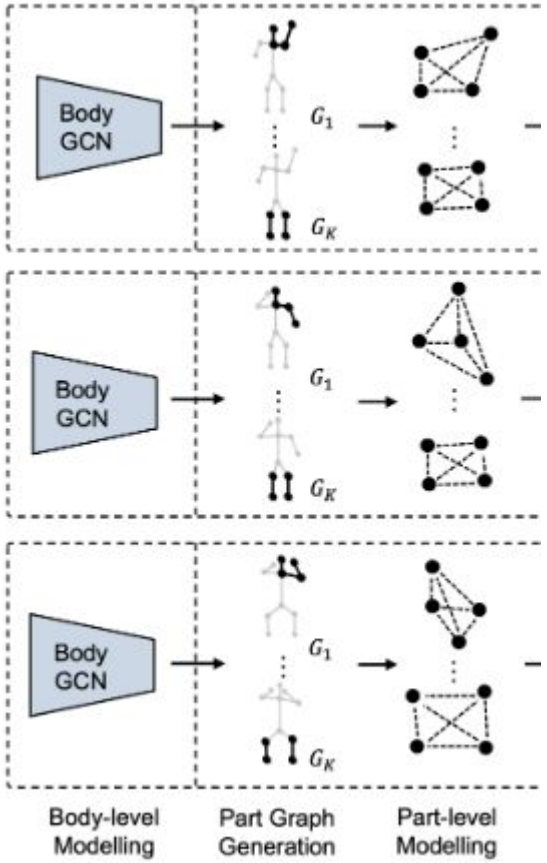
- Objetivo: $Distance = D(\mathbf{x}^q, \mathbf{x}^s)$
- *Multi-part embeding* $\{\Gamma_1, \Gamma_2, \ldots, \Gamma_K\} = \mathscr{F}_{embed}(\mathbf{x})$
- Fuse embedings $\boldsymbol{\varepsilon} = \mathscr{F}_{fuse}(\Gamma_1, \Gamma_2, \ldots, \Gamma_K)$

Metodologia: visão geral da Arquitetura



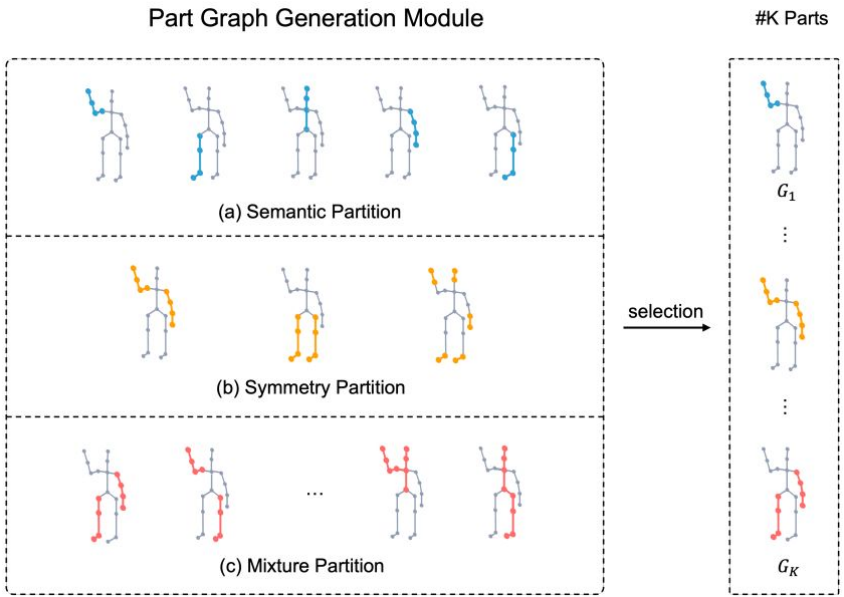
Arquitetura: *Cascaded Embedding Module.*

$$\{\Gamma_1, \Gamma_2, \ldots, \Gamma_K\} = \mathscr{F}_{embed}(\mathbf{x})$$



Arquitetura: *Cascaded Embedding Module.*

Foi gerado poses com as juntas usando como base um conjunto de regras:
a. partição semântica
b. partição simétrica
c. mistura com as partições semântica e simétrica.



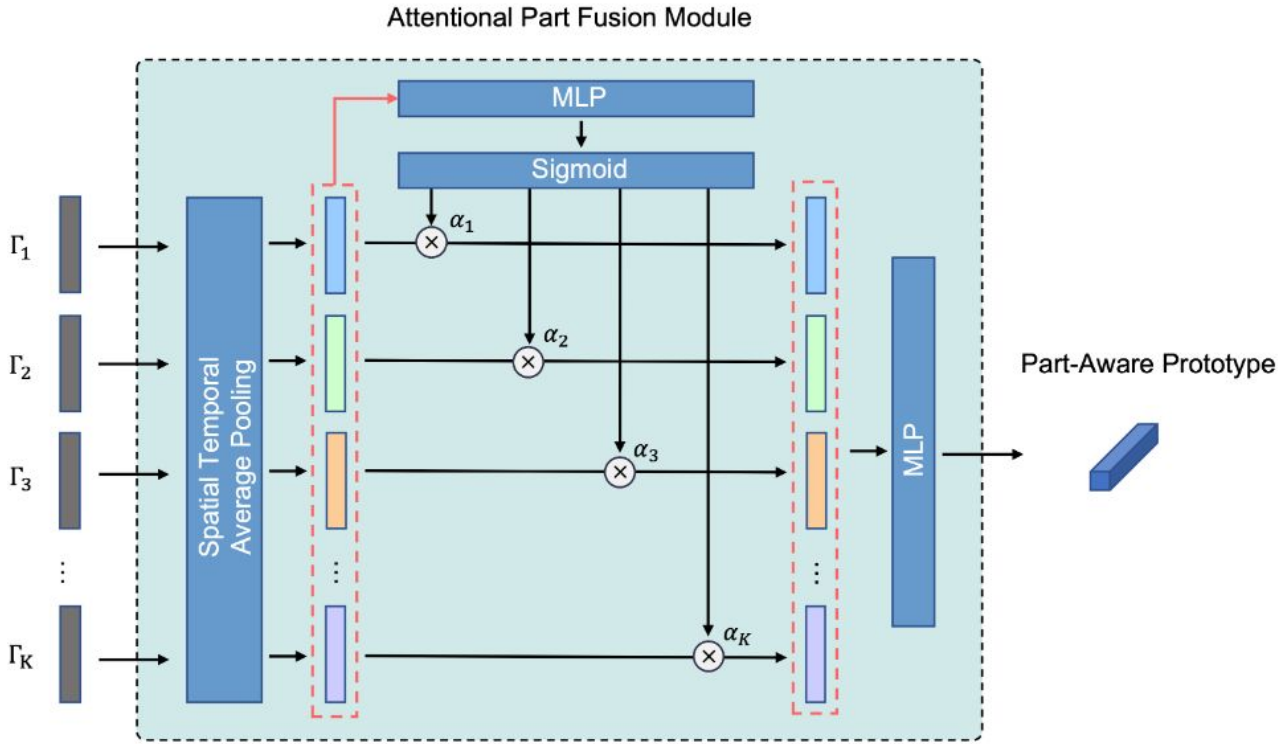
Particionamento de poses.

Arquitetura: *Attention part fusion.*

$$\varepsilon = \mathrm{MLP}(\gamma_1' \oplus \gamma_2' \oplus \ldots \oplus \gamma_K').$$

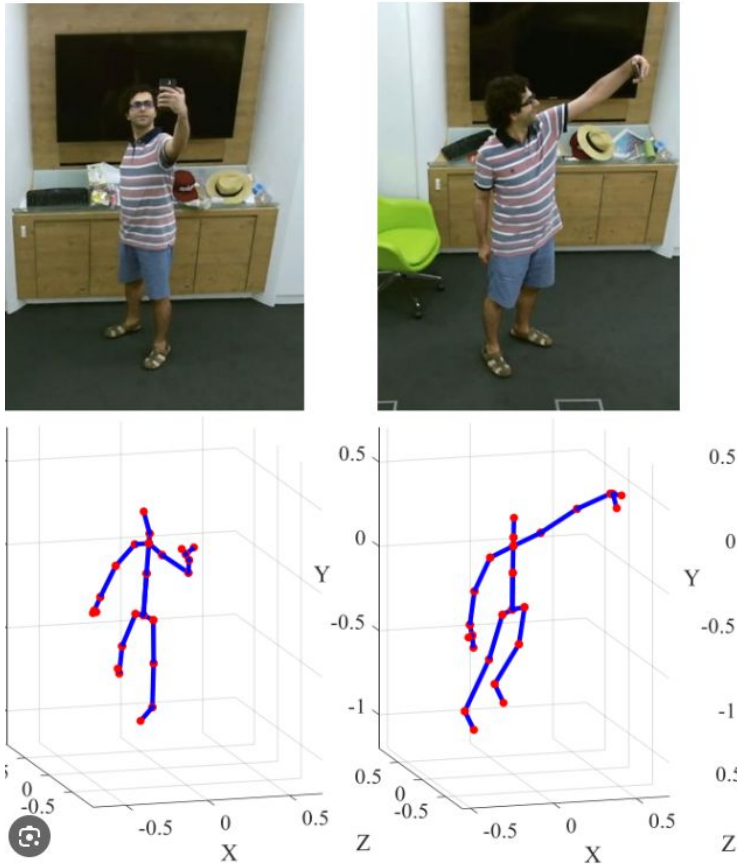
$$d(\varepsilon^q, \varepsilon^s) = -(\frac{\varepsilon^q}{||\varepsilon^q||})^T \cdot \frac{\varepsilon^s}{||\varepsilon^s||}.$$

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$



Bases de dados experimentadas.

- NTU RGB + D 120.
  - 120 classes.
- NW-UCLA.
  - 10 classes.



Resultados: NTU RGB+D 120.

# Training Classes	20	40	60	80	100
APSR [10]	29.1	34.8	39.2	42.8	45.3
SL-DML [15]	36.7	42.4	49.0	46.4	50.9
Skeleton-DML [14]	28.6	37.5	48.6	48.0	54.2
JEANIE [25]	38.5	44.1	50.3	51.2	57.0
ProtoNet [21]+ST-GCN [27]	41.5	49.6	54.2	55.2	61.1
ProtoNet [21]+MV-IGNet [26]	41.6	49.2	53.1	54.5	60.1
ProtoNet [21]+MS-G3D [13]	41.1	48.7	54.4	52.7	59.5
ProtoNet [21]+CTR-GCN [3]	39.9	49.1	53.6	54.2	58.8
Ours	**43.0**	**50.3**	**55.7**	**56.5**	**65.6**

Resultados: NW-UCLA.

Method	Accuracy(%)
SL-DML [15] †	65.6
Skeleton-DML [14] †	72.8
ProtoNet [21]+ST-GCN [27]	79.8
ProtoNet [21]+MV-IGNet [26]	80.9
ProtoNet [21]+MS-G3D [13]	81.2
ProtoNet [21]+CTR-GCN [3]	80.7
Ours	**83.3**

Resultados: ablação. *Self-attention heads*.

# Heads	Params	Accuracy(%)
1	1.9M	**65.6**
2	2.2M	61.3
4	2.8M	60.6
8	3.9M	59.6

Resultados: ablação. Modelos de atenção.

Method	Accuracy(%)
w/o attention	62.9
Self-attention	61.2
MLP-attention (Ours)	**65.6**

Conclusão.

- Forma interessante de registar espaço de juntas locais.
- Conceitos de cascata também interessantes aplicados diretamente ao problema.
- Mostrando resultados, estado da arte no reconhecimento de ações *one-shot* .

Dúvidas, sugestões e discussão.

- Será que colocar uma camada de atenção *joint-wise* traria melhorias?
- Será que o método é bom para ações que a pose sofre oclusão?
  - E ainda mais, onde as juntas importantes para ação sofrem oclusão?
- Adicionar um *dropout* de conjunto de juntas, no treino da rede em multi-shot, traria melhorias em few-shot e multi-shot para cenas de poses com oclusão?