Improving Action Recognition with Ordered Skeletal Joints

Olayinka Ajayi, Hongkai Wen & Tanaya Guha

1. Problem of Graph-based Action Recognition

Many of the action recognition algorithm use a dynamic graph to obtain spatial embeddings. This increases the tendency that some nodes might have the same embedding (see Fig 1), and this can lead to a drop in the recognition accuracy of the model. This problem is closely related to the semantics of a sentence depending on the position of a word in the sentence. In NLP, this was resolved by encoding the position of the word in the embeddings [1].

Unlike sequences (or grids) that are naturally ordered, graphs have no inherent order. We propose an algorithm to learn an ordering for the nodes of the human skeletal graph.

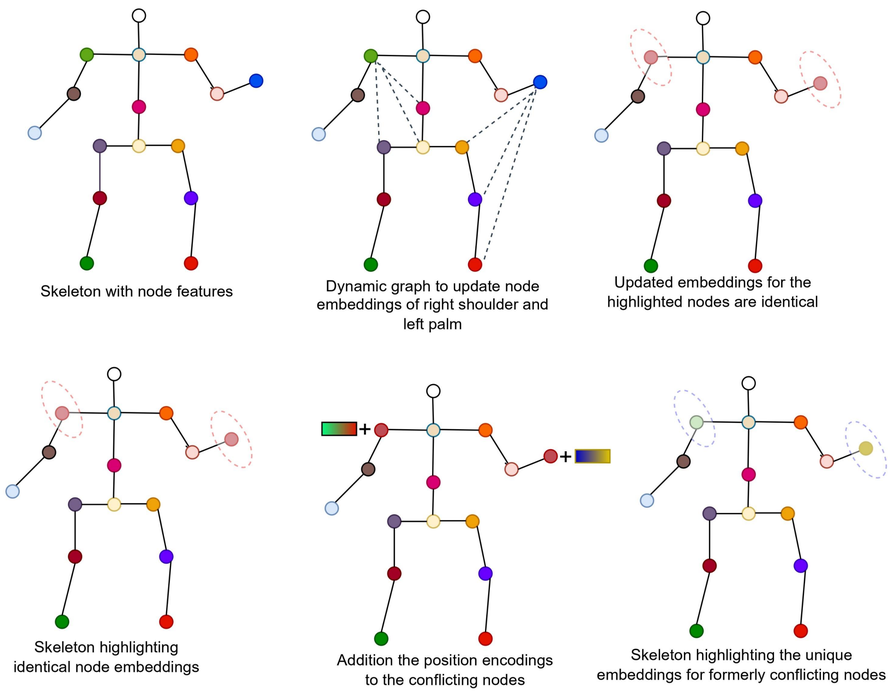

Fig 1

2. Method

A part of our method adopts a variation of the supervised contrastive learning proposed by Khosla et al. [2] to cluster similar embeddings. We use the hamming distance $d(.)$ in place of the dot product. Our method learns a binary encoding $\boldsymbol{Z}_i$ for each node using the equation below (see Fig 2).

$$\mathcal{L} = \mathcal{L}_{adj} + k_1\mathcal{L}_{deg\_dist} + k_2\mathcal{L}_{deg},$$

$$\mathcal{L}_{deg\_dist} := \sum_{i \in V} -\frac{1}{|\deg(i)|} \sum_{\rho \in \deg(i)} \log \frac{\exp\left(d(\mathbf{Z}_i, \mathbf{Z}_\rho)\right)}{\sum_{a \in \kappa(i)} \exp\left(d(\mathbf{Z}_i, \mathbf{Z}_a)\right)},$$

$$\mathcal{L}_{adj} := \sum_{i \in V} -\frac{1}{|\mathcal{N}(i)|} \sum_{\rho \in \mathcal{N}(i)} \log \frac{\exp\left(d(\mathbf{Z}_i, \mathbf{Z}_\rho)\right)}{\sum_{a \in \mu(i)} \exp\left(d(\mathbf{Z}_i, \mathbf{Z}_a)\right)},$$

$$\mathcal{L}_{deg} := ||\mathbf{Z}\mathbf{W}_{deg} - \mathbf{D}||_2$$

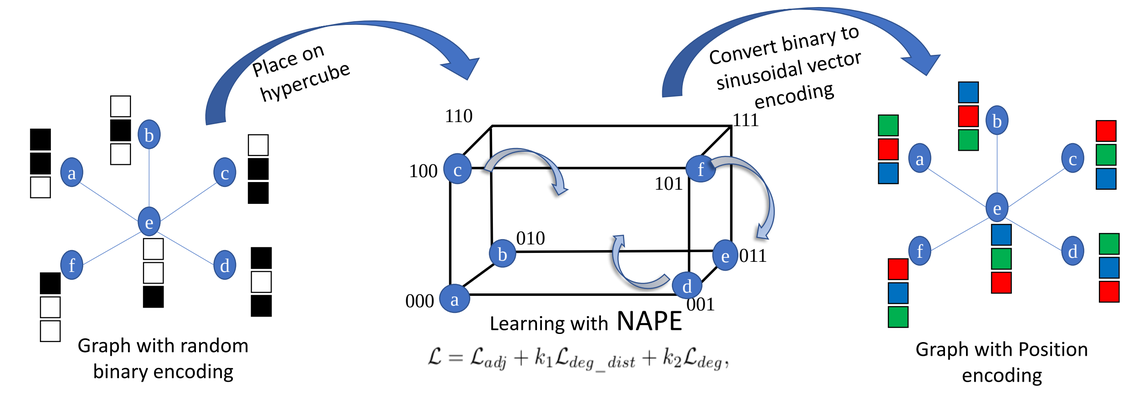

Fig 2

3. Numbering As a Position Encoding (NAPE)

The final binary encodings are converted to decimals (see Fig 3). These decimals are converted to sinusoidal vectors using the equation proposed in the Attention is all you need paper [1]. Our method produces a systematic ordering of the nodes (in red) which we call Numbering As a Position Encoding (NAPE).

$$\overrightarrow{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k.t), & \text{if } i = 2k \\ \cos(\omega_k.t), & \text{if } i = 2k+1 \end{cases} \qquad \omega_k = \frac{1}{10000^{2k/d}}$$

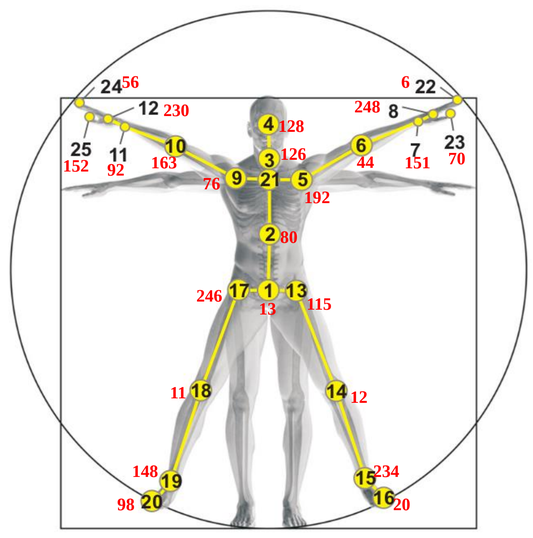

Fig 3

4. Results

We compare the accuracy of our action recognition model with and without the NAPE position encoding. We also observe that using NAPE, the InfoGCN [3] action recognition model saves 3200 FLOPs for each epoch during training without sacrificing recognition accuracy.

Recognition Algorithm	w/o PE	NAPE	1-25 PE	Learned PE	FLOPS Saved
Our Method	89.1%	**91.0%**	89.7%	90.2%	-
InfoGCN [3]	-	**95.4%**	-	**95.4%**	3200

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need.

[2] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... & Krishnan, D. (2020). Supervised contrastive learning.

[3] Chi, H. G., Ha, M. H., Chi, S., Lee, S. W., Huang, Q., & Ramani, K. (2022). Infogcn: Representation learning for human skeleton-based action recognition. In Proceedings of the IEEE/CVF.

Contact

+44 7493 929215
Olayinka.Ajayi@warwick.ac.uk

Mathematics for Real-world Systems CDT

SCAN ME