Spatial Temporal Transformer Network for Skeleton-based Action Recognition

Chiara Plizzari, Marco Cannici, and Matteo Matteucci

Politecnico di Milano, Milano, Italy
chiara.plizzari@mail.polimi.it
{marco.cannici, matteo.matteucci}@polimi.it

Abstract. Skeleton-based human action recognition has achieved a great interest in recent years, as skeleton data has been demonstrated to be robust to illumination changes, body scales, dynamic camera views, and complex background. Nevertheless, an effective encoding of the latent information underlying the 3D skeleton is still an open problem. In this work, we propose a novel Spatial-Temporal Transformer network (ST-TR) which models dependencies between joints using the Transformer self-attention operator. In our ST-TR model, a Spatial Self-Attention module (SSA) is used to understand intra-frame interactions between different body parts, and a Temporal Self-Attention module (TSA) to model inter-frame correlations. The two are combined in a two-stream network which outperforms state-of-the-art models using the same input data on both NTU-RGB+D 60 and NTU-RGB+D 120.

1 Introduction

Skeleton-based activity recognition is achieving increasing interest in recent years thanks to advances in 3D skeleton pose estimation devices, both in terms of accuracy and resolution. Algorithms and neural architectures for extracting context-aware fine-grained spatial-temporal features, capable of unlocking the true potential of skeleton based action recognition, however, are still lacking in the literature. The most widespread method to perform skeleton-based action recognition has become Spatial-Temporal Graph Convolutional Network (ST-GCN) [16], since, being an efficient representation of non-Euclidean data, it is able to effectively capture spatial (intra-frame) and temporal (inter-frame) information. However, ST-GCN models have some structural limitations, some of them already addressed in [13,14,2,10]: (i) The topology of the graph representing the human body is fixed for all layers and all the actions, preventing the extraction of rich representations. (ii) Both the spatial and temporal convolutions are implemented from a standard 2D convolution. As such, they are limited to operate in a local neighborhood. (iii) As a consequence of (i) and (ii), correlations between body joints not linked in the human skeleton, e.g., the left and right hands, are underestimated even if relevant in actions such as "clapping".

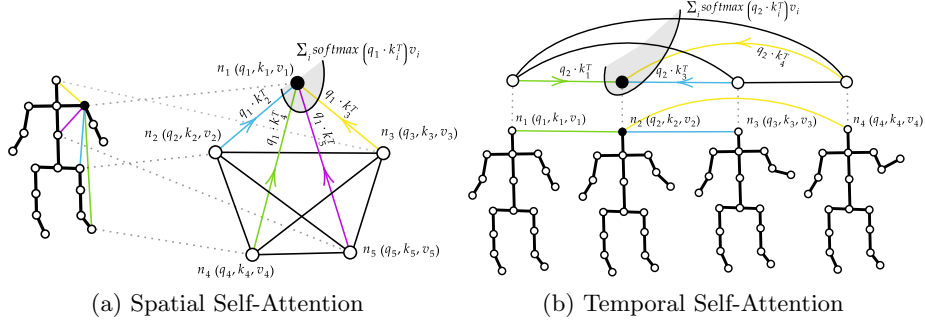(a) Spatial Self-Attention   (b) Temporal Self-Attention

Fig. 1. Spatial Self-Attention (SSA) and Temporal Self-Attention (TSA). Self-attention operates on each pair of nodes by computing a weight for each of them representing the strength of their correlation. Those weights are used to score the contribution of each body joint n_i, proportionally to how relevant the node is w.r.t. the other ones.

In this paper, we face all these limitations by employing a modified Transformer self-attention operator. Despite being originally designed for Natural Language Processing (NLP) tasks, the flexibility of the Transformer self-attention [15] in modeling long-range dependencies, make this model a perfect solution to tackle ST-GCN weaknesses. Recently, Bello et al. in [1] employed self-attention on image pixels to overcome the locality of the convolution operator. In our work, we aim to apply the same mechanism to joints representing the human skeleton, with the goal of extracting adaptive low-level features modeling interactions in human actions both in space, through a Spatial Self-Attention module (SSA), and time, through a Temporal Self-Attention module (TSA) module. Authors of [3] also proposed a Self-Attention Network (SAN) to extract long-term semantic information; however, since they focus on temporally segmented clips, they solve the locality limitations of convolution only partially.

Contributions of this paper are summarized as follows:

- We propose a novel two-stream Transformer-based model, employing *self-attention* on both the spatial and temporal dimensions
- We design a *Spatial Self-Attention* (SSA) module to dynamically build links between skeleton joints, representing the relationships between human body parts, conditionally on the action and independently from the natural human body structure. On the temporal dimension, we introduce a *Temporal Self-Attention* (TSA) module to study the dynamics of joints along time too [1]
- Our model outperforms the ST-GCN [16] baseline, and outperforms previous state-of-the-art methods using the same input data on NTU-RGB+D [12,6].

---

[1] Code at https://github.com/Chiaraplizz/ST-TR

2   Spatial Temporal Transformer Network

We propose *Spatial Temporal Transformer (ST-TR)*, an architecture using the Transformer self-attention mechanism to operate both on space and time. We develop two modules, *Spatial Self-Attention (SSA)* and *Temporal Self-Attention (TSA)*, each one focusing on extracting correlations in one of the two dimensions.

2.1   Spatial Self-Attention (SSA)

The SSA module applies self-attention *inside each frame* to extract low-level features embedding the relations between body parts, i.e., computing correlations between each pair of joints in every single frame independently, as depicted in Figure 1a. Given the frame at time t, for each node i^t of the skeleton, a *query* vector $\mathbf{q}_i^t \in \mathbb{R}^{dq}$, a *key* vector $\mathbf{k}_i^t \in \mathbb{R}^{dk}$ and a *value* vector $\mathbf{v}_i^t \in \mathbb{R}^{dv}$ are first computed by applying trainable linear transformations to the node features $\mathbf{n}_i^t \in \mathbb{R}^{C_{in}}$, shared across all nodes, of parameters $\mathbf{W}_q \in \mathbb{R}^{C_{in} \times dq}$, $\mathbf{W}_k \in \mathbb{R}^{C_{in} \times dk}$, $\mathbf{W}_v \in \mathbb{R}^{C_{in} \times dv}$. Then, for each pair of body nodes (i^t, j^t), a *query-key dot product* is applied to obtain a weight $\alpha_{ij}^t \in \mathbb{R}$ representing the strength of the correlations between the two nodes. The resulting score α_{ij}^t is used to weight each joint value $\mathbf{v}_j^t$, and a weighted sum is computed to obtain a new embedding for node i^t, as in the following:

$$\alpha_{ij}^t = \mathbf{q}_i^t \cdot \mathbf{k}_j^{t\,T}, \forall t \in T, \qquad \mathbf{z}_i^t = \sum_j softmax_j \left(\frac{\alpha_{ij}^t}{\sqrt{d_\mathrm{k}}} \right) \mathbf{v}_j^t \qquad (1)$$

where $\mathbf{z}_i^t \in \mathbb{R}^{C_{out}}$ (with C_{out} the number of output channels) constitutes the new embedding of node i^t.

Multi-head self-attention is finally applied by repeating this embedding extraction process H times, each time with a different set of learnable parameters. The set $(\mathbf{z}_{i_1}^t, ..., \mathbf{z}_{i_H}^t)$ of node embeddings thus obtained, all referring to the same node i^t, is finally combined with a learnable transformation, i.e., $concat(\mathbf{z}_{i_1}^t, ..., \mathbf{z}_{i_H}^t) \cdot \mathbf{W}_o$, and constitutes the output features of SSA.

Thus, as shown in Figure 1a, the relations between nodes (i.e., the α_{ij}^t scores) are dynamically *predicted* in SSA; the correlation structure is not fixed for all the actions, but it changes adaptively for each sample. SSA operates similar to a graph convolution on a fully connected graph where, however, the kernel values (i.e., the α_{ij}^t scores) are predicted dynamically based on the skeleton pose.

2.2   Temporal Self-Attention (TSA)

Along the temporal dimension, the dynamics of each joint is studied separately *along all the frames*, i.e., each single node is considered as independent and correlations between frames are computed by comparing features of the same body joint along the temporal dimension (see Figure 1b). The formulation is

symmetrical to the one reported in Equation (1) for SSA:

$$\alpha_{ij}^v = \mathbf{q}_i^v \cdot \mathbf{k}_j^v \quad \forall v \in V, \qquad \mathbf{z}_i^v = \sum_j softmax_j \left(\frac{\alpha_{ij}^v}{\sqrt{d_\mathrm{k}}} \right) \mathbf{v}_j^v \qquad (2)$$

where i^v, j^v indicate the same joint v in two different instants i, j, $\alpha_{ij}^v \in \mathbb{R}$, $\mathbf{q}_i^v \in \mathbb{R}^{dq}$ is the query associated to i^v, $\mathbf{k}_j^v \in \mathbb{R}^{dk}$ and $\mathbf{v}_j^v \in \mathbb{R}^{dv}$ are the key and value associated to joint j^v (all computed using trainable linear transformations as in SSA), and $\mathbf{z}_i^v \in \mathbb{R}^{C_{out}}$ is the resulting node embedding. An illustration of TSA is depicted in Figure 1b. Multi-head attention is applied as in SSA. The network, by extracting inter-frame relations between nodes in time, can learn to correlate frames apart from each other (e.g., nodes in the first frame with those in the last one), capturing discriminant features that are not otherwise possible to capture with a standard convolution, being this limited by the kernel size.

2.3   Two-Stream Spatial Temporal Transformer Network

To combine the SSA and TSA modules, a two-stream architecture named 2s-ST-TR is used, as similarly proposed by Shi et al. in [14] and [13]. In our formulation, the two streams differentiate on the way the self-attention mechanism is applied: SSA operates on the spatial stream (named S-TR stream), while TSA on the temporal one (named T-TR stream). On both streams, simple features are first extracted through a three-layers residual network, where each layer processes the input on the spatial dimension through graph convolution (GCN), and on the temporal dimension through a standard 2D convolution (TCN), as in ST-GCN [16]. SSA and TSA are applied on the S-TR and on the T-TR stream in substitution to the GCN and TCN feature extraction modules respectively (Figure 2). Each stream is trained using the standard cross-entropy loss, and the sub-networks outputs are eventually fused together by summing up their softmax output scores to obtain the final prediction, as in [14,13].

Spatial Transformer Stream (S-TR). In the spatial stream (Figure 2), self-attention is applied at the skeleton level through a SSA module, which focuses on spatial relations between joints, and then its output is passed to a 2D convolutional module with kernel K_t on temporal dimension (TCN), as in [16], to extract temporally relevant features, i.e., $\textbf{S-TR}(x) = Conv_{2D(1 \times K_t)}(\textbf{SSA}(x))$. Following the original Transformer structure, the input is pre-normalized passing through a Batch Normalization layer [4,11], and skip connections are used, which sum the input to the output of the SSA module.

Temporal Transformer Stream (T-TR). The temporal stream, instead, focuses on discovering inter-frame temporal relations. Similarly to the S-TR stream, inside each T-TR layer, a standard graph convolution sub-module [16] is followed by the proposed Temporal Self-Attention module, i.e., $\textbf{T-TR}(x) = \textbf{TSA}(GCN(x))$. In this case, TSA operates on graphs linking the same joint along all the time dimension (e.g., all left feet, or all right hands).
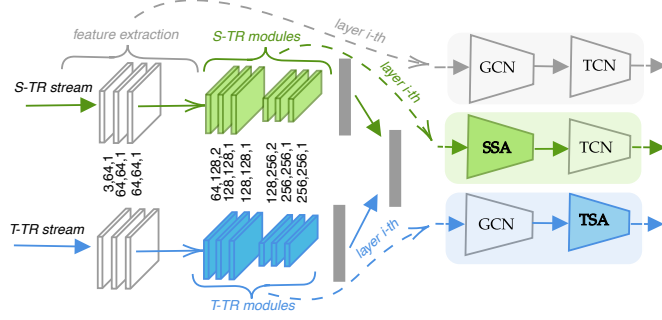
Fig. 2. The 2s-ST-TR architecture. On each stream, the first three layers extract simple features. On the S-TR stream, at each subsequent layer, SSA is used to extract spatial information, followed by a 2D convolution on time dimension (TCN). On the T-TR stream, at each subsequent layer, TSA is used to extract temporal information, while spatial features are extracted by a standard graph convolution (GCN) [16]

3 Model Evaluation

3.1 Datasets

NTU RGB+D 60 and NTU RGB+D 120. The NTU RGB+D 60 (NTU-60) dataset is a large-scale benchmark for 3D human action recognition [12]. Skeleton information consists of 3D coordinates of 25 body joints and a total of 60 different action classes. The NTU-60 dataset follows two different criteria for evaluation. In *Cross-View Evaluation* (X-View), the data is split according to the camera from which the action is taken, while in *Cross-Subject Evaluation* (X-Sub) according to the subject performing the action. NTU-RGB+D 120 [6] (NTU-120) is an extension of NTU-60, with a total of $113,945$ videos and 120 classes. It follows two evaluation criteria: *Cross-Subject Evaluation (X-Sub)* is the same used in NTU-60, while in *Cross-Setup Evaluation (X-Set)* training and testing samples are split based on the parity of the camera setup IDs.

3.2 Experimental Settings

Using PyTorch framework, we trained our models for a total of 120 epochs with batch size 32 and SGD as optimizer. The learning rate is set to 0.1 at the beginning and then reduced by a factor of 10 at the epochs {60, 90}. Moreover, we preprocessed the data with the same procedure used by Shi et al. in [14] and [13]. In order to avoid overfitting, we also used *DropAttention* [17], a dropout technique introduced by Zehui et al. [17] for regularizing attention weights in Transformer. In all of these experiments, the *number of heads* for multi-head attention is set to 8, and d_q, d_k, d_v embedding dimensions to $0.25 \times C_{out}$ in each layer, as in [1]. We did not perform a grid search on these parameters.

(a)	Method	Bones	X-Sub	X-View
	ST-GCN [16]		85.7	92.7
	S-TR		86.4	94.0
	T-TR		86.0	93.6
	ST-TR		88.7	95.6
	S-TR	✓	87.9	94.9
	T-TR	✓	87.3	94.1
	ST-TR	✓	89.9	96.1

(b)	Module	Params [$\times 10^4$]
	GCN [16]	19.9
	SSA	17.8
	TCN [16]	59.0
	TSA	17.7

Table 1. a) Accuracy (%) comparison of S-TR and T-TR, and their combination (ST-TR) on NTU-60, w and w/o bones. b) Parameters of SSA and TSA modules

3.3   Results

To verify the effectiveness of our SSA and TSA modules, we compare separately the S-TR stream and T-TR stream against the ST-GCN [16] baseline, whose results on NTU-60 [12] are reported in Table 1 using our learning rate scheduling.

As far as it concerns the SSA, S-TR outperforms the baseline by 0.7% on X-Sub, and by 1.3% on X-View, demonstrating that self-attention can be used in place of graph convolution, increasing the network performance while also decreasing the number of parameters. On NTU-60 the S-TR stream achieves slightly better performance (+0.4%) than the T-TR stream, on both X-View and X-Sub (Table 1a). This can be motivated by the fact that SSA in S-TR operates on 25 joints only, while on the temporal dimension the number of correlations is proportional to the huge number of frames. Table 1b shows the difference in terms of parameters between a single GCN (TCN) and the corresponding SSA (TSA) module, with $C_{in} = C_{out} = 256$. Especially on the temporal dimension, TSA results in a decrease in parameters, introducing 41.3×10^4 less than TCN. The combination of the two streams achieves 88.7% of accuracy on X-Sub and 95.6% of accuracy on X-View, outperforming the baseline ST-GCN by up to 3% and surpassing other two-stream architectures (Table 2). Classes that benefit the most from self-attention are "playing with phone", "typing", and "cross hands" on S-TR, and those involving long-range relations or two people, i.e., "hugging", "point finger", "pat on back", on T-TR. These require to correlate along the entire action, giving empirical insight on the advantage of the proposed method.

As adding bones information demonstrated leading to better results in previous works [13,14], we also studied the effect of our self-attention module on combined joint and bones information. For each node $\mathbf{v}_1 = (\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1)$ and $\mathbf{v}_2 = (\mathbf{x}_2, \mathbf{y}_2, \mathbf{z}_2)$, the bone connecting the two is calculated as $\mathbf{b}_{\mathbf{v}_1, \mathbf{v}_2} = (\mathbf{x}_2 - \mathbf{x}_1, \mathbf{y}_2 - \mathbf{y}_1, \mathbf{z}_2 - \mathbf{z}_1)$. Joint and bone information are concatenated along the channel dimension and then fed to the network. At each layer, the size of the input and output channels is doubled as in [13,14]. The performance results are shown again in Table 1a; all previous configurations improve when bones are added as input. The latter fact highlights the flexibility of our method, which is capable of adapting to different input types and network configurations.

(a)

NTU-60			
Method	Bones	X-Sub	X-View
ST-GCN [16]		81.5	88.3
1s-AGCN [14][10]		86.0	93.7
1s Shift-GCN [2]		87.8	95.1
SAN [3]		87.2	92.7
ST-TR (Ours)		**88.7**	**95.6**
2s-AGCN [14]	✓	88.5	95.1
DGCNN [13]	✓	89.9	96.1
2s Shift-GCN [2]	✓	89.7	96.0
MS-G3D [10]	✓	**91.5**	**96.2**
ST-TR (Ours)	✓	89.9	96.1

(b)

NTU-120		
Method	X-Sub	X-Set
ST-LSTM [7]	55.7	57.9
GCA-LSTM [8]	61.2	63.3
RotClips+MTCNN [5]	62.2	61.8
Pose Evol. Map [9]	64.6	66.9
1s Shift-GCN [2]	80.9	83.2
S-TR (Ours)	78.6	80.7
T-TR (Ours)	78.4	80.5
ST-TR (Ours)	**81.9**	**84.1**

Table 2. Comparison with state-of-the-art accuracy (%) of S-TR, T-TR, and their combination (ST-TR) on NTU-60 (a) and NTU-120 (b)

4   Comparison with State-Of-The-Art

We compare our methods on NTU-60 and NTU-120 w.r.t. other methods which make use of joint or joint+bones information on a one- or two-stream architecture, as we also did, for a fair comparison (Table 2). On NTU-60, ST-TR without bones outperforms all the state-of-the-art models not using bones, including 1s-AGCN and SAN [3], which uses self-attention too. Similarly, our ST-TR with bones outperforms all previous two-stream methods that use bones information as well, i.e., 2s-AGCN and 2s Shift-GCN. On NTU-120, the model based only on joint information outperforms all state-of-the-art methods making use of the same information. The competitive results validate the superiority of our method over architectures relying on convolution only.

5   Conclusions

In this paper, we propose a novel approach that introduces Transformer self-attention in skeleton activity recognition as an alternative to graph convolution. Through experiments on NTU-60 and NTU-120, we demonstrated that our Spatial Self-Attention module (SSA) can replace graph convolution, enabling more flexible and dynamic representations. Similarly, the Temporal Self-Attention module (TSA) overcomes the strict locality of standard convolution, leading to global motion pattern extraction. Moreover, our final Spatial-Temporal Transformer network (ST-TR) achieves state-of-the-art performance on NTU-RGB+D w.r.t. methods using same input information and streams setup.

References

1. Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V.: Attention augmented convolutional networks. Proceedings of the IEEE International Conference on Computer

Vision pp. 3286–3295 (2019)

2. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 183–192 (2020)

3. Cho, S., Maqbool, M., Liu, F., Foroosh, H.: Self-attention network for skeleton-based human action recognition. In: The IEEE Winter Conference on Applications of Computer Vision. pp. 635–644 (2020)

4. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)

5. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: Learning clip representations for skeleton-based 3d action recognition. IEEE Transactions on Image Processing **27**(6), 2842–2855 (2018)

6. Liu, J., Shahroudy, A., Perez, M.L., Wang, G., Duan, L.Y., Chichung, A.K.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence (2019)

7. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. European conference on computer vision pp. 816–833 (2016)

8. Liu, J., Wang, G., Duan, L.Y., Abdiyeva, K., Kot, A.C.: Skeleton-based human action recognition with global context-aware attention lstm networks. IEEE Transactions on Image Processing **27**(4), 1586–1599 (2017)

9. Liu, M., Yuan, J.: Recognizing human actions as the evolution of pose estimation maps. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 1159–1168 (2018)

10. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 143–152 (2020)

11. Nguyen, T.Q., Salazar, J.: Transformers without tears: Improving the normalization of self-attention. arXiv preprint arXiv:1910.05895 (2019)

12. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 1010–1019 (2016)

13. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 7912–7921 (2019)

14. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 12026–12035 (2019)

15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems pp. 5998–6008 (2017)

16. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. Thirty-second AAAI conference on artificial intelligence (2018)

17. Zehui, L., Liu, P., Huang, L., Fu, J., Chen, J., Qiu, X., Huang, X.: Dropattention: A regularization method for fully-connected self-attention networks. arXiv preprint arXiv:1907.11065 (2019)