

On the spatial attention in Spatio-Temporal Graph Convolutional Networks for skeleton-based human action recognition

Negar Heidari and Alexandros Iosifidis

Department of Electrical and Computer Engineering, Aarhus University, Denmark

Emails: {negar.heidari, alexandros.iosifidis}@eng.au.dk

Abstract—Graph convolutional networks (GCNs) achieved promising performance in skeleton-based human action recognition by modeling a sequence of skeletons as a spatio-temporal graph. Most of the recently proposed GCN-based methods improve the performance by learning the graph structure at each layer of the network using a spatial attention applied on a predefined graph Adjacency matrix that is optimized jointly with model's parameters in an end-to-end manner. In this paper, we analyze the spatial attention used in spatio-temporal GCN layers and propose a symmetric spatial attention for better reflecting the symmetric property of the relative positions of the human body joints when executing actions. We also highlight the connection of spatio-temporal GCN layers employing additive spatial attention to bilinear layers, and we propose the spatio-temporal bilinear network (ST-BLN) which does not require the use of predefined Adjacency matrices and allows for more flexible design of the model. Experimental results show that the three models lead to effectively the same performance. Moreover, by exploiting the flexibility provided by the proposed ST-BLN, one can increase the efficiency of the model.

Index Terms—Graph Convolutional Network, Bilinear Network, Spatial Attention, Human Action Recognition, Spatio-Temporal Graph

I. INTRODUCTION

Considering the availability of depth cameras and successful pose estimation toolboxes such as OpenPose [1], each action can be represented by a sequence of body poses, each of which can be represented by a skeleton. Compared to other data modalities which are used for human action recognition, such as RGB videos, depth images and optical flow, human body skeleton represents the body pose and motion in a compact graph structure which is robust to context noise and invariant to body scale, view point variations, lighting conditions and background context [2]. Thus, skeleton-based human action recognition has become of great interest in recent years. Many of the recently proposed methods either use skeletons data in conjunction with other data modalities, such as RGB images [3], or only utilize the skeleton data to efficiently extract high level features for human action recognition [4]–[6].

Many deep learning methods have been proposed recently for skeleton-based human action recognition which are mainly categorized into Recurrent Neural Network (RNN)-based, Convolutional Neural Network (CNN)-based, and Graph Convolutional Network (GCN)-based methods. RNN-based methods are the earliest methods proposed in this field which

mostly utilize Long Short-Term Memory (LSTM) networks [7] to model the temporal dynamics of the sequence of skeletons [8]–[13]. These methods represent each skeleton in a sequence as a vector which is formed by concatenated human body joints' coordinates. The CNN-based methods [14]–[19] employ the state-of-the-art CNN methods to extract the spatial and temporal features of the sequence of skeletons and they represent each skeleton in a sequence as a pseudo-image which is formed by reorganizing the body joints' coordinates into a 2D matrix. Since RNN-based and CNN-based methods convert the skeleton data, which has an irregular structure, into a regular sequence or grid structure, it cannot benefit from the non-Euclidean structure of the skeleton sequences.

Recently, several GCN-based methods have been proposed for skeleton-based human action recognition and they achieved state-of-the-art performance [20]–[23] by utilizing the graph structure of the skeleton data. In skeleton-based human action recognition, the temporal dynamics of each action are represented by a sequence of skeletons and each human body skeleton is modeled as a graph which encodes the spatial structure of human body joints and their natural connections. Spatio-temporal GCN (ST-GCN) method [20] is the first GCN-based method proposed for skeleton-based human action recognition which receives a sequence of skeletons as input and employs GCN layers to capture both spatial and temporal features in actions by exploiting the graph structure of input data. The performance of ST-GCN method has been improved by several methods which build on top of ST-GCN. These methods mostly try to adaptively learn the graph structure in each GCN layer in an end-to-end manner, based on the features of input data [21]–[23]. 2s-AGCN [22] learns the graph structure adaptively based on the graph joints' similarity in the input data and also the existing physical connections in the body. GCN-NAS [23] is a neural architecture search method which explores the search space with different graph modules to improve the representational capacity of the GCN layers. Moreover, the attention mechanism has been employed in these methods in order to highlight the most important connections in body skeleton for each action class. AS-GCN [24] extended the skeleton graphs to represent both structural links and actional linked and proposed an actional-structural graph convolutional network, which has an encoder-decoder structure, to capture richer dependencies from actions. DPRL+GCNN [25] and TA-GCN [26] select the most informative skeletons in a sequence

to make the inference process more efficient.

In this paper, we analyze the attention mechanisms of existing spatio-temporal GCN networks used for human action recognition. Based on this analysis, we make two contributions. First, we argue that the attention mechanism in GCN layers using an additive formulation should lead to a symmetric attention mask, as the learned attentions corresponding to a pair of nodes should reflect the symmetric relationship of the corresponding human body joint positions in a human body pose. We propose a symmetric attention mechanism that is shown to perform on par with the original one for different structures of the model. Second, we propose a spatio-temporal bilinear layer which allows for more flexible design of the model for skeleton-based human action recognition. We show that models with spatio-temporal bilinear layers perform on par with spatio-temporal GCN networks without requiring the design of graph structures to encode relationships of the joints of the body skeletons.

The remainder of the paper is organized as follows. Section II introduces the ST-GCN method which is the background of our work. Section III describes the spatial attention used in GCN-based human action recognition and Section IV analyzes its properties and proposes an alternative symmetric spatial attention. Section V describes the spatio-temporal bilinear network, which is motivated by the analysis of the spatial attention in GCN-based human action recognition methods. The experimental results are provided in section VI and the conclusions are drawn in section VII.

II. SPATIO-TEMPORAL GCN (ST-GCN)

In this section, the ST-GCN method [20] is introduced as the baseline of many recent GCN-based methods for skeleton-based human action recognition. This method receives a sequence of skeletons $\mathbf{X} \in \mathbb{R}^{C^{in} \times T \times V}$ encoding the human body poses comprising an action as input, where C^{in} is the number of input channels, T is the number of skeletons in the sequence, and V is the number of joints in each skeleton. It then models the sequence of skeletons as a spatio-temporal graph and applies multiple GCN layers with spatial and temporal convolutions to extract high level features for classifying the input skeleton sequence to a set of pre-defined action classes. A spatio-temporal graph on a sequence of skeletons is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} is the set of the human body joints and \mathcal{E} denotes both spatial and temporal edges connecting the body joints within each skeleton and between skeletons. The spatial edges are the natural physical connections between human body joints and the temporal edges connect each body joint of a skeleton to its corresponding joint of previous and subsequent skeletons. Therefore, the spatial graph is constructed based on the human body structure in which each node can have different number of neighbors, while in the temporal graph each node has two fixed neighbors, i.e. the nodes corresponding to the same joint in the previous and next time step. Figure 1 (right) shows the spatio-temporal graph on a sequence of skeletons.

Since the body motions can be grouped to concentric and eccentric, ST-GCN employs a spatial partitioning process to

divide the neighboring set of each node into three subsets based on their spatial arrangement in a single skeleton. This partitioning process fixes the node degrees in the spatial graph and defines the partitions as: 1) the root node itself, 2) the root node's neighbors which are closer to the skeleton's center of gravity than the root node and 3) the remaining root node's neighbors that are farther from the skeleton's center of gravity. The skeleton center of gravity (COG) is denoted as the average of all skeleton joints' coordinates. In Figure 1 (left) a spatially partitioned graph is illustrated in which the nodes in different neighboring subsets (partitions) are shown with different colors and the center of gravity is shown as a red dot.

According to the partitioning process, the graph structure in each skeleton is captured by three binary Adjacency matrices, which are hereafter indexed by p , each of which encodes the structure of a neighboring node subset. The Adjacency matrix of the first partition (corresponding to the root nodes) indicates the nodes' self-connections and is set to $\mathbf{A}_1 = \mathbf{I}_V$. Each adjacency matrix is normalized as $\hat{\mathbf{A}}_p = \mathbf{D}_p^{-\frac{1}{2}} \mathbf{A}_p \mathbf{D}_p^{-\frac{1}{2}}$, where $\mathbf{D}^{(ii)}_p = \sum_j \mathbf{A}_{(ij)}_p + \varepsilon$ denotes the degree matrix and $\varepsilon = 0.001$ is used to avoid division with zero that can be caused by empty rows in $\mathbf{D}^{(ii)}_p$. The element $\hat{\mathbf{A}}_{p,ij}$ indicates whether the vertex j is in the p^{th} neighboring subset of vertex i .

Each GCN layer updates the human body features through both spatial and temporal domains by applying spatial and temporal convolutions on the output of the previous layer. The spatial convolution is defined based on the layer-wise propagation rule of GCNs [27] and the l^{th} layer spatial convolution is given by:

$$\mathbf{H}_s^{(l)} = ReLU \left(\sum_p \left(\hat{\mathbf{A}}_p \otimes \mathbf{M}_p^{(l)} \right) \mathbf{H}^{(l-1)} \mathbf{W}_p^{(l)} \right), \quad (1)$$

where \otimes is the element-wise product of two matrices and $\mathbf{H}^{(l-1)} \in \mathbb{R}^{C^{(l-1)} \times T^{(l-1)} \times V}$ denotes the output of the $(l-1)^{th}$ layer which is introduced to the l^{th} layer as input. The input of the first layer is defined as $\mathbf{H}^{(0)} = \mathbf{X}$. Since there are three spatial graphs for each sample, the GCN's propagation rule is applied on each graph with a different weight matrix $\mathbf{W}_p^{(l)} \in \mathbb{R}^{C^{(l)} \times C^{(l-1)}}$, where $C^{(l-1)}$ and $C^{(l)}$ denote the number of channels in layers $l-1$ and l , respectively. $\mathbf{M}_p^{(l)} \in \mathbb{R}^{V \times V}$ is a learnable attention matrix which highlights the most important connections in a skeleton for each action. It is initialized as an all-one matrix and is element-wise multiplied to its corresponding adjacency matrix of each graph partition. The spatial convolution is in practice a 2D convolution operation which performs $C^{(l)}$ convolutions with filters of size $C^{(l-1)} \times 1 \times 1$ on the input tensor and the resulting output is a tensor of size $C^{(l)} \times T^{(l-1)} \times V$ which is multiplied with the masked attention-based adjacency matrix $\hat{\mathbf{A}}_p \otimes \mathbf{M}_p^{(l)}$ on the last dimension V .

To benefit from the motions taking place in each action for label prediction, the output of the spatial convolution $\mathbf{H}_s^{(l)}$ is introduced to the temporal convolution block which is also a 2D convolution that applies $C^{(l)}$ convolutions with filters of size $C^{(l)} \times k \times 1$ to capture the temporal dynamics

in the sequence of skeletons by propagating the information through the time domain while keeping the number of channels unchanged. Based on the filter size k , the temporal convolution propagates the features through k consecutive skeletons and reduces the number of skeletons or keeps it unchanged depending on the stride and padding size. Accordingly, the output of temporal convolution is of size $C^{(l)} \times T^{(l)} \times V$ in which $T^{(l)}$ denotes the temporal dimension of sequence in layer l . In ST-GCN method, the kernel size for temporal convolution in each layer is set to $k = 9$. The extracted spatio-temporal features of the last ST-GCN layer are introduced to a fully connected classification layer which is equipped by a global average pooling block and a SoftMax activation function. The entire model is trained in an end-to-end manner using Backpropagation.

III. THE ROLE OF SPATIAL ATTENTION IN GCN-BASED HUMAN ACTION RECOGNITION

One of the main drawbacks of ST-GCN method which is addressed by more recently proposed methods, such as 2s-AGCN [22], is that it uses a fixed graph structure which is heuristically predefined and represents the natural physical connections between the body joints in a skeleton. The attention mechanism in ST-GCN, by using an element-wise multiplication between the Adjacency matrix and the learnable attention mask, can only highlight or diminish existing connections between the body joints which is not guaranteed to be optimal for action classification task. For some actions such as “touching head”, the connection between the hand and head is important for recognizing the action, while this connection does not exist naturally in the body and, thus, is not considered in the predefined graph structure. Therefore, 2s-AGCN defines the spatial convolution as follows:

$$\mathbf{H}_s^{(l)} = \text{ReLU} \left(\sum_p \left(\hat{\mathbf{A}}_p + \mathbf{M}_p^{(l)} \right) \mathbf{H}^{(l-1)} \mathbf{W}_p^{(l)} \right), \quad (2)$$

where $\mathbf{M}_p^{(l)}$ is a learnable attention matrix added to the graph in order to both highlight/diminish existing connections between the skeleton joints and also add potentially important connections between the disconnected ones. This matrix is initialized by zeros and it is learned in an end-to-end manner along with the other model’s parameters. In other words, the spatial graph is only initialized by the skeleton’s structure, and it is updated adaptively by an attention matrix whose parameters are learned in the training process. Comparing the attention-based Adjacency matrices of ST-GCN and 2s-AGCN, $\hat{\mathbf{A}}_p \otimes \mathbf{M}_p^{(l)}$ and $\hat{\mathbf{A}}_p + \mathbf{M}_p^{(l)}$ respectively, it can be seen that the former is guaranteed to have a structure encoding the natural connections between the human body joints, while the latter corresponds to a matrix encoding a fully-connected graph structure.

IV. SYMMETRIC SPATIAL ATTENTION FOR GCN-BASED HUMAN ACTION RECOGNITION

We focus on the properties of the attention-based Adjacency matrix $\hat{\mathbf{A}}_p + \mathbf{M}_p^{(l)}$, as it has been shown to improve action

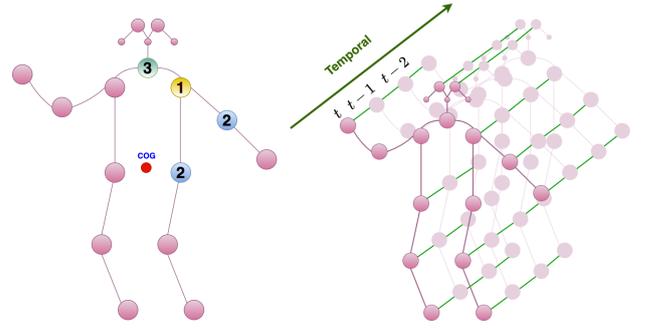


Fig. 1: Illustration of an example Spatio-temporal graph (right), and the neighboring subsets in spatial partitioning process in different colors (left).

classification performance compared to the one given by $\hat{\mathbf{A}}_p \otimes \mathbf{M}_p^{(l)}$. It can be shown that our analysis also holds for the latter case. Since the matrix $\mathbf{M}_p^{(l)}$ is optimized in an end-to-end manner jointly with the parameters of the entire network without imposing any constraints, as detailed in Eq. (2), its final form will be an asymmetric matrix containing both positive and negative values. This form of the attention mask can be qualitatively explained by considering that the attention put by a graph node v_i to another graph node v_j can be freely optimized and can differ from the attention put by v_j to v_i . However, considering that the nodes of the graph correspond to human body skeleton joints and during action execution their relative positions are symmetric, we would expect that their pair-wise attentions should be the same. In order to enforce this property in the optimization process of the attention-based Adjacency matrix we define the attention mask to be a symmetric matrix, i.e. $\mathbf{M}_p^{(l)} = \mathbf{L}_p^{(l)} \mathbf{L}_p^{(l)T}$. Accordingly, we define the spatial convolution as follows:

$$\mathbf{H}_s^{(l)} = \text{ReLU} \left(\sum_p \left(\hat{\mathbf{A}}_p + \mathbf{L}_p^{(l)} \mathbf{L}_p^{(l)T} \right) \mathbf{H}^{(l-1)} \mathbf{W}_p^{(l)} \right). \quad (3)$$

While one can select $\mathbf{M}_p^{(l)}$ to be of low rank by selecting $\mathbf{L}_p^{(l)} \in \mathbb{R}^{V \times q}$ with $q < V$, we do not set any further restrictions to $\mathbf{M}_p^{(l)}$, except of being symmetric. Once the training process ends, the matrix $\mathbf{M}_p^{(l)}$ is calculated and used for inference, thus, the space and time complexities remain the same as in the case of using Eq. (2). Here we should note that, since the normalized Adjacency matrices $\hat{\mathbf{A}}_p$ are designed to be asymmetric, the final attention-based Adjacency matrices will be asymmetric too.

V. SPATIO-TEMPORAL BILINEAR NETWORK

Considering the optimization of the attention-based Adjacency matrices based on the definition of the spatial convolution of ST-GCN in Eqs. (2) and (3), it can be seen that the addition of the normalized Adjacency matrix $\hat{\mathbf{A}}_p$ to the attention mask serves as an initialization strategy, while the final form of the attention-based Adjacency matrix depends primarily on the learned values of the mask. This means that, after the first few training epochs, the GCN blocks of the ST-GCN using an additive attention mask do not follow the graph

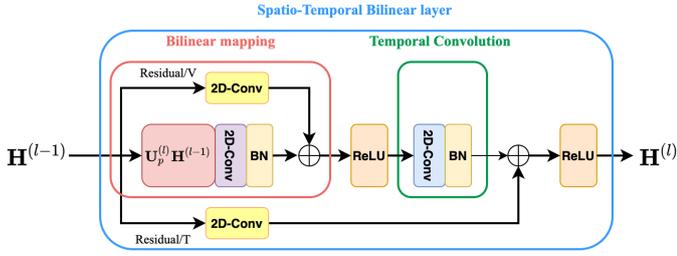


Fig. 2: Illustration of spatio-temporal bilinear layer l receiving as input $\mathbf{H}^{(l-1)}$ of size $C^{(l-1)} \times T^{(l-1)} \times V^{(l-1)}$ providing an output $\mathbf{H}^{(l)}$ of size $C^{(l)} \times T^{(l)} \times V^{(l)}$. 2D-Conv blocks correspond to standard 2D convolutions. The 2D-Conv block in bilinear mapping transforms the data by applying $C^{(l)}$ convolutions with filters of size $C^{(l-1)} \times 1 \times 1$ to change the number of channels. The 2D-Conv block in temporal convolution keeps the number of channels unchanged and applies $C^{(l)}$ filters of size $C^{(l)} \times k \times 1$ in order to aggregate the features in temporal domain. The residual connections also apply 2D convolutions on the layer’s input to have the same dimension as the layer’s output. Residual/V indicates the residual connection which adds the input to the output of bilinear mapping, and Residual/T denotes the residual connection which adds the layer’s input to the output of temporal convolution. BN and ReLU represent the batch-normalization and ReLU activation function, respectively.

structure anymore, but they are rather equivalent to bilinear layers [28]–[30]. That is, Eqs. (2) and (3) can take the form:

$$\mathbf{H}_s^{(l)} = \text{ReLU} \left(\sum_p \mathbf{U}_p^{(l)} \mathbf{H}^{(l-1)} \mathbf{W}_p^{(l)} \right), \quad (4)$$

where $\mathbf{U}_p^{(l)} \in \mathbb{R}^{V^{(l)} \times V^{(l-1)}}$ is a learnable matrix which is optimized in an end-to-end manner jointly with the parameters of the entire network. A specific setting of the matrix $\mathbf{U}_p^{(l)}$ for which $V^{(l-1)} = V^{(l)} = V$ corresponds to the attention-based Adjacency matrix in Eqs. (2) and (3) receiving as input the representations for the V human body joints at layer l and transforming them by applying a data transformation using $\mathbf{W}_p^{(l)}$ and combining information of neighboring nodes according to the connectivity in $\mathbf{U}_p^{(l)}$. However, by using the bilinear layer definition in Eq. (4) one can freely decide on the dimensions of the transformation to be applied using $\mathbf{U}_p^{(l)}$. That is, considering that in the first layer of the Spatio-Temporal Bilinear network the matrix $\mathbf{U}_p^{(1)} \in \mathbb{R}^{V^{(1)} \times V}$ performs a transformation in the mode of the input $\mathbf{H}^{(0)}$ corresponding to the human body skeleton joints, selection of $V^{(1)} < V$ will lead to aggregation of joints’ information to create a new set of $V^{(1)}$ nodes. On the other hand, selection of $V^{(1)} > V$ will lead to the creation of new nodes to be processed by the second layer. A similar explanation can be given to the selection of the values $V^{(l)}$ for layer l . A special case in this setting is the one where $V^{(l)} = 1$, leading to the creation of one node encoding information of the entire input skeleton data.

Similar to ST-GCN, a spatio-temporal bilinear layer is

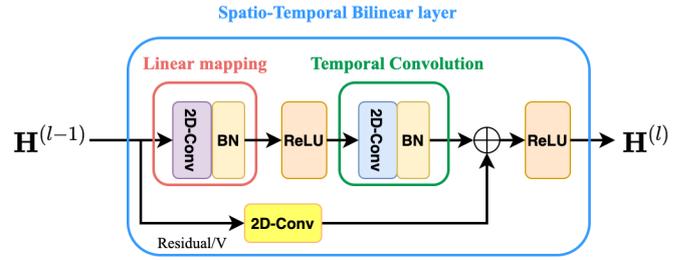


Fig. 3: Illustration of spatio-temporal bilinear layer l when $V^{(l-1)} = V^{(l)} = 1$.

equipped with a batch normalization, residual connections and ReLU activation function. The structure of a bilinear layer is shown in Figure. 2. The layer has two residual connections to stabilize the model by summing up the layer’s input with the output of bilinear mapping and the second one adds the layer’s input to the output of temporal convolution. We can distinguish two cases for the residual connections. When $V^{(l-1)} = V^{(l)}$, the number of dimensions of the input tensor $\mathbf{H}^{(l-1)}$ to layer l changes by applying the transformation in one of its modes using $\mathbf{W}_p^{(l)}$, and the residual connections need to perform a 2D convolution with $C^{(l)}$ filters of size $C^{(l-1)} \times 1 \times 1$ on the layer’s input to have the same channel dimension as the layer’s output $\mathbf{H}^{(l)}$. When $V^{(l-1)} \neq V^{(l)}$, the number of dimensions of the input tensor $\mathbf{H}^{(l-1)}$ to layer l changes by applying the transformations in both of its modes using $\mathbf{W}_p^{(l)}$ and $\mathbf{U}_p^{(l)}$, and the residual connections need to perform a 2D convolution with $C^{(l)} \cdot V^{(l)}$ filters of size $C^{(l-1)} \times 1 \times V^{(l-1)}$ on the layer’s input. The resulting tensor is then reshaped to a tensor of $C^{(l)} \times T^{(l)} \times V^{(l)}$ to have the same dimension as the layer’s output $\mathbf{H}^{(l)}$.

If at layer $l - 1$ a value of $V^{(l-1)} = 1$ is used, the input $\mathbf{H}^{(l-1)}$ to layer l becomes of size $C^{(l-1)} \times T^{(l-1)} \times 1$. In that case there is no need of using $\mathbf{U}_p^{(l)} \in \mathbb{R}$ as the scaling factor that would be learned by using it can be absorbed in $\mathbf{W}_p^{(l)}$. Thus, the bilinear mapping can be replaced with an equivalent linear mapping. Moreover, in this case there is no need of using a residual connection for the linear mapping block. The architecture of such a spatio-temporal bilinear layer is shown in Figure 3.

Similar to the ST-GCN method, the output of the last spatio-temporal bilinear layer is passed to a global average pooling layer to leading to 256-dimensional features for each sequence and the resulting vector is introduced to a fully connected classification layer which is equipped with a SoftMax activation function. The entire model is trained in an end-to-end manner using Backpropagation to optimize the objective function defined on the output using the classification loss.

VI. EXPERIMENTS

We conducted two sets of experiments on NTU-RGB+D dataset [10] which is the largest indoor-captured action recognition dataset and it is widely used for evaluating the skeleton-based human action recognition methods. It consists of 56, 880 action video clips from 60 different human action classes and

Attention/Mapping	CS(%)	CV(%)
$\hat{\mathbf{A}}_p + \mathbf{M}_p^{(l)}$	85.29	93.78
$\hat{\mathbf{A}}_p + \mathbf{L}_p^{(l)} \mathbf{L}_p^{(l)T}$	87.14	93.86
$\mathbf{U}_p^{(l)}$	85.71	93.80

TABLE I: Comparison of classification accuracy of spatio-temporal networks equipped with asymmetric and symmetric spatial attentions, and the bilinear mappings on the test set of CV benchmark of NTU-RGB+D dataset. The models are trained using joints data.

#Layers	$\hat{\mathbf{A}}_p + \mathbf{M}_p^{(l)}$	$\hat{\mathbf{A}}_p + \mathbf{L}_p^{(l)} \mathbf{L}_p^{(l)T}$	$\mathbf{U}_p^{(l)}$
1	77.09	77.16	78.39
2	87.53	87.00	88.14
3	89.90	89.93	89.13
4	90.58	90.80	90.51
5	91.96	92.22	92.15
6	92.89	93.15	92.40
7	93.37	93.37	93.51
8	93.51	93.95	93.74
9	93.73	93.76	93.52
10	93.78	93.86	93.80

TABLE II: Comparison of classification accuracy of spatio-temporal networks equipped with asymmetric and symmetric spatial attentions, and the bilinear mappings on the test set of CV benchmark of NTU-RGB+D dataset for different number of layers. The models use the same layer sizes as those in [22] and are trained using joints data.

each clip is captured by 3 cameras from 3 different views. Each action clip is used to extract a sequence of 300 skeletons, each of which is represented by the 3D coordinates of 25 body joints. This leads to each action video clip being represented as a $3 \times 300 \times 25$ tensor. Two benchmarks are defined in [10], i.e. cross view (CV) and cross-subject (CS), and a train-test split for each benchmark is also provided. We use the same experimental protocols and data splits as in [10]. The CV benchmark contains 37,920 training samples captured by cameras two and three and 18,960 test samples captured by the first camera. In CS benchmark, the training set contains 40,320 samples and test set contains 16,560 samples while the actors in training and test set are different.

The experiments are conducted on PyTorch deep learning framework [31] with 4 GRX 1080-ti GPUs. We used the same experimental settings and network architectures as in 2s-AGCN [22]. The model is trained for 50 epochs with SGD optimizer and cross entropy loss function. The mini-batch size is set to 64 and the learning rate is initialized to 0.1 and it is divided by 10 at epochs 30 and 40.

In the first set of experiments, we compared the performance of spatio-temporal networks using GCN layers equipped with attentions in Eqs. (2) and (3) and using bilinear layer as defined in Eq. (4). The performance of the three models is reported in Table I. As can be seen, the three networks achieved very similar performance. This indicates that there is no idfference in using the predefined graph structure encoded by the matrices $\hat{\mathbf{A}}_p$ in Eqs. (2) and (3) for initializing the mapping in the bilinear layer (4). To further analyze the effect of the models'

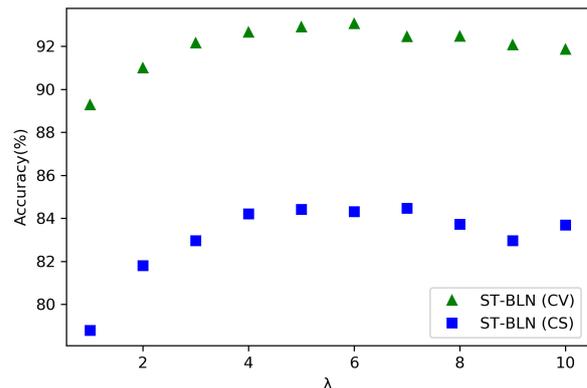


Fig. 4: Performance of ST-BLN model in terms of classification accuracy, when joints' features are aggregated in different layers λ of the network.

size and the differences in performance achieved by the three types of spatio-temporal networks, we conducted multiple experiments using different number of layers and the results are reported in Table II. As can be seen, the performance of the three types of networks for all network sizes are effectively the same. This confirms our observations related to the role of the predefined graph structure used in the spatial attention of spatio-temporal networks for skeleton-based action recognition. We can conclude that the combining information related to the nodes of the human body skeleton graph can be randomly initialized and optimized jointly with the other network parameters, and that the use of predefined skeleton graph structures is not necessary for recognizing the human actions using skeleton data.

For completeness, we also provide in Table III the performance of state-of-the-art methods in skeleton-based human action recognition on the two benchmarks of the NTU-RGB+D dataset. Since most of the state-of-the-art methods, like 2s-AGCN, GCN-NAS, AS-GCN and 2s-TA-GCN, train their model with two or more data streams, we also report the performance achieved by the spation-temporal bilinear network using two streams. That is, the 2s-ST-BLN model is formed by two streams, the first of which receives as input the joints data and the second one receives as input the bone data. The two streams process their inputs and the predicted SoftMax scores of the two streams are fused to obtain the final classification outcome. The results show that ST-BLN and 2s-ST-BLN perform on par with other state-of-the-art methods while they do not require a predefined graph structure to encode the relationships between the joints (and bones) of the human body skeleton. Similar to the GCN-based methods, the proposed method also outperforms the RNN-based and CNN-based methods with a large margin.

To evaluate the need of combining information in the dimension of human body skeleton joints, expressed either as spatial attention of the form in Eq. (2) and as bilinear mapping of the form in Eq. (4), or if it is adequate the employ linear mappings effectively leading to neural layers combining

Method	CS(%)	CV(%)	#Streams
HBRNN [8]	59.1	64.0	5
Deep LSTM [10]	60.7	67.3	1
ST-LSTM [9]	69.2	77.7	1
STA-LSTM [11]	73.4	81.2	1
VA-LSTM [12]	79.2	87.7	1
ARRN-LSTM [13]	80.7	88.8	2
2s-3DCNN [14]	66.8	72.6	2
TCN [15]	74.3	83.1	1
Clips+CNN+MTLN [16]	79.6	84.8	1
Synthesized CNN [17]	80.0	87.2	1
3scale ResNet152 [18]	85.0	92.3	1
CNN+Motion+Trans [19]	83.2	89.3	2
ST-GCN [20]	81.5	88.3	1
DPRL+GCNN [25]	83.5	89.8	1
TA-GCN [26]	87.97	94.2	1
AS-GCN [24]	86.8	94.2	2
2s-AGCN [22]	88.5	95.1	2
2s-TA-GCN [26]	88.5	95.1	2
GCN-NAS [23]	89.4	95.7	2
ST-BLN	85.71	93.80	1
2s-ST-BLN	87.8	95.1	2

TABLE III: Classification accuracy comparison of spatio-temporal bilinear model with state-of-the-art GCN-based methods on the test set of NTU-RGB+D dataset.

a Multilayer Perceptron block with a Temporal Convolution block, we conducted a second set of experiments. In this set of experiments, we used a 10-layer spatio-temporal bilinear network formed by the same data transformation sizes as in the first set of our experiments, but instead of using bilinear mappings with $V^{(l)} = V$ for all 10 layers of the model, we use $V^{(l)} = V$, $l = 1, \dots, \lambda - 1$ and $V^{(l)} = 1$, $l = \lambda, \dots, 10$. That is, the first $\lambda - 1$ layers of the model use Spatio-Temporal Bilinear layers as the one illustrated in Figure 2 with values $V^{(l)} = V$. At layer λ a Spatio-Temporal Bilinear layer in the form of Figure 2 is used with $V^{(\lambda)} = 1$, while the remaining layers take the form of a Spatio-Temporal Bilinear layer as the one illustrated in Figure 3. We applied experiments using both the CV and CS benchmarks of the NTU-RGB+D dataset using the joints data, and the results in terms of classification accuracy and number of FLOPs are indicated in Figures 4 and 5, respectively. As can be seen in Figure 4, on the CV benchmark the best performance is equal to 93.06% and is obtained for a value of $\lambda = 6$. On the CS benchmark, the best performance is obtained for a value of $\lambda = 7$ and is equal to 84.47%. These performance values are on par with the results reported in Table I, achieved by using the Spatio-Temporal Bilinear/GCN layers using $V^{(l)} = V$, $l = 1, \dots, 10$. Observing the performance values achieved by the different networks in Figure 4, we can conclude that combining information in the dimension of human body skeleton joints at early layers of the network contributes in the performance achieved by the network, while this is not the case for deeper layers in the network.

By observing the number of Floating Point Operations needed to evaluate an input skeleton sequence for different choices of λ in Figure 5, we can see that one can increase

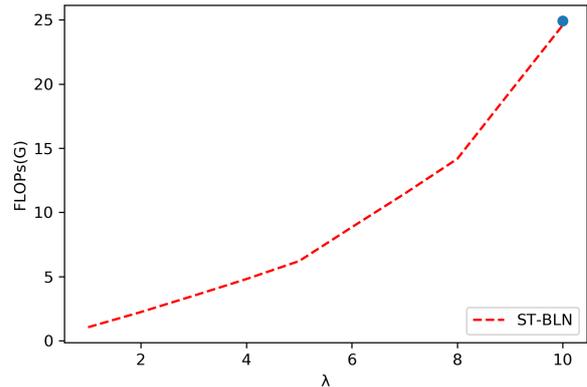


Fig. 5: Floating Point Operations of the ST-BLN model when joints' features are aggregated in different layers λ of the network. The case of a ST-BLN with $V^{(l)} = V$, $l = 1, \dots, 10$ is illustrated with a dot.

efficiency of the model by using lower values of λ . This is due to that for layers beyond the λ -th layer of the network a smaller number of parameters is needed, as indicated by comparing the two layers illustrated in Figures 2 and 3. Considering that the number of model's parameters at the deeper layers is higher compared to the first layers, it can be seen that a considerable improvement in the model's efficiency can be achieved. Overall, a network using a value of $\lambda = 6$ ($\lambda = 7$) operates $\times 2.78$ ($\times 2.14$) faster compared to a network with $\lambda = 10$. As expected, the numbers of FLOPs for a network with $\lambda = 10$ and a ST-BLN with all layers using values $V^{(l)} = V$ are very similar, i.e. 24.59G and 24.93G, respectively.

VII. CONCLUSION

In this paper, we studied the properties of the spatial attention used in ST-GCN methods for skeleton-based human action recognition. We made two observations, the first being that the spatial attention used in these methods leads to an asymmetric attention matrix. Our second observation is that the human-expert designed normalized Adjacency matrices used in models with additive attention serve only as an initialization process, and do not really affect parameters of the network for combining information in the dimension of human body skeleton joints used for inference. Based on the first observation, we proposed a symmetric spatial attention that can be directly incorporated in the existing ST-GCN methods without affecting the overall number of computations during inference. Based on the second observation, we proposed the Spatio-Temporal Bilinear Network (ST-BLN) which does not require the use of a predefined Adjacency matrix and allows for more flexible design of the model for skeleton-based human action recognition. Extensive experimental analysis shows that the three models lead to effectively the same performance. Moreover, by exploiting the flexibility provided by the proposed ST-BLN, one can increase the model's efficiency

by a factor of more than $\times 2$ while preserving the performance levels of the original method.

ACKNOWLEDGEMENT

This work was supported by the European Union’s Horizon 2020 Research and Innovation Action Program under Grant 871449 (OpenDR). This publication reflects the authors’ views only. The European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

- [1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [2] F. Han, B. Reily, W. Hoff, and H. Zhang, “Space-time representation of people based on 3D skeletal data: A review,” *Computer Vision and Image Understanding*, vol. 158, pp. 85–105, 2017.
- [3] A. Franco, A. Magnani, and D. Maio, “A multimodal approach for human activity recognition based on skeleton and rgb data,” *Pattern Recognition Letters*, vol. 131, pp. 293–299, 2020.
- [4] D. C. Luvizon, H. Tabia, and D. Picard, “Learning features combination for human action recognition from skeleton sequences,” *Pattern Recognition Letters*, vol. 99, pp. 13–20, 2017.
- [5] A. Saggese, N. Strisciuglio, M. Vento, and N. Petkov, “Learning skeleton representations for human action recognition,” *Pattern Recognition Letters*, vol. 118, pp. 23–31, 2019.
- [6] M. Li and H. Leung, “Graph-based approach for 3d human skeletal action recognition,” *Pattern Recognition Letters*, vol. 87, pp. 195–202, 2017.
- [7] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A search space odyssey,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2016.
- [8] Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [9] J. Liu, A. Shahroudy, D. Xu, and G. Wang, “Spatio-temporal LSTM with trust gates for 3D human action recognition,” in *European Conference on Computer Vision*, 2016.
- [10] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “NTU RGB+D: A large scale dataset for 3D human activity analysis,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [11] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, “An end-to-end spatio-temporal attention model for human action recognition from skeleton data,” in *AAAI Conference on Artificial Intelligence*, 2017.
- [12] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, “View adaptive recurrent neural networks for high performance human action recognition from skeleton data,” in *IEEE International Conference on Computer Vision*, 2017.
- [13] L. Li, W. Zheng, Z. Zhang, Y. Huang, and L. Wang, “Relational network for skeleton-based action recognition,” in *IEEE International Conference on Multimedia & Expo*, 2019.
- [14] H. Liu, J. Tu, and M. Liu, “Two-stream 3d convolutional neural network for skeleton-based action recognition,” *arXiv preprint arXiv:1705.08106*, 2017.
- [15] T. S. Kim and A. Reiter, “Interpretable 3D human action analysis with temporal convolutional networks,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.
- [16] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, “A new representation of skeleton sequences for 3D action recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [17] M. Liu, H. Liu, and C. Chen, “Enhanced skeleton visualization for view invariant human action recognition,” *Pattern Recognition*, vol. 68, pp. 346–362, 2017.
- [18] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, “Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN,” in *IEEE International Conference on Multimedia & Expo Workshops*, 2017.
- [19] C. Li, Q. Zhong, D. Xie, and S. Pu, “Skeleton-based action recognition with convolutional neural networks,” in *IEEE International Conference on Multimedia & Expo Workshops*, 2017.
- [20] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *AAAI Conference on Artificial Intelligence*, 2018.
- [21] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Skeleton-based action recognition with directed graph neural networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [22] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [23] W. Peng, X. Hong, H. Chen, and G. Zhao, “Learning graph convolutional network for skeleton-based human action recognition by neural searching,” in *AAAI Conference on Artificial Intelligence*, 2020.
- [24] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, “Action-structural graph convolutional networks for skeleton-based action recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [25] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, “Deep progressive reinforcement learning for skeleton-based action recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [26] N. Heidari and A. Iosifidis, “Temporal attention-augmented graph convolutional network for efficient skeleton-based human action recognition,” in *International Conference on Pattern Recognition*, 2020.
- [27] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *International Conference on Learning Representations*, 2017.
- [28] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, “Compact bilinear pooling,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [29] D. T. Tran, A. Iosifidis, J. Kannianen, and M. Gabbouj, “Temporal attention-augmented bilinear network for financial time-series data analysis,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1407–1418, 2018.
- [30] J.-H. Kim, J. Jun, and B.-T. Zhang, “Bilinear attention networks,” in *Advances in Neural Information Processing Systems*, 2018.
- [31] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” in *Advances in Neural Information Processing Systems Workshops*, 2017.