electronics

MDPI

Article

2s-GATCN: Two-Stream Graph Attentional Convolutional Networks for Skeleton-Based Action Recognition

Shu-Bo Zhou, Ran-Ran Chen, Xue-Qin Jiang * and Feng Pan

Institute of Information Science and Technology, Donghua University, Shanghai 201620, China; zhoushubo@dhu.edu.cn (S.-B.Z.)
* Correspondence: xqjiang@dhu.edu.cn

Abstract: As human actions can be characterized by the trajectories of skeleton joints, skeleton-based action recognition techniques have gained increasing attention in the field of intelligent recognition and behavior analysis. With the emergence of large datasets, graph convolutional network (GCN) approaches have been widely applied for skeleton-based action recognition and have achieved remarkable performances. In this paper, a novel GCN-based approach is proposed by introducing a convolutional block attention module (CBAM)-based graph attention block to compute the semantic correlations between any two vertices. By considering semantic correlations, our model can effectively identify the most discriminative vertex connections associated with specific actions, even when the two vertices are physically unconnected. Experimental results demonstrate that the proposed model is effective and outperforms existing methods.

Keywords: action recognition; GCN; connection strength; graph attention block; CBAM

Citation: Zhou , S.-B.; Chen, R.-R.; Jiang, X.-Q.; Pan, F. 2s-GATCN: Two-Stream Graph Attentional Convolutional Networks for Skeleton-Based Action Recognition. *Electronics* **2023**, *12*, 1711. https://doi.org/10.3390/electronics12071711

Academic Editor: Donghyeon Cho

Received: 14 March 2023
Revised: 1 April 2023
Accepted: 2 April 2023
Published: 4 April 2023

1. Introduction

As an important research problem in the field of intelligent recognition and behavior analysis, human action recognition methods [1] have gained significant importance in various fields, such as video surveillance, virtual reality, smart homes, three-dimensional perceptions, and human–computer interactions [2–5]. These methods can be categorized into two types: vision-based methods [6,7] and skeleton-based methods [8,9]. Vision-based methods recognize human actions by extracting spatial or temporal contextual features from image sequences or videos, while skeleton-based methods represent the human body as a three-dimensional graph, and extract features from vertices and edges to recognize actions. In comparison to vision-based data, such as RGB images and videos, 3D skeleton data is less computationally expensive and more robust in complex background. Thus, this paper focuses on exploring 3D skeleton-based methods for human action recognition.

The DL-based action recognition methods can be typically categorized into recurrent neural network (RNN)-based methods [10,11], convolutional neural network (CNN)-based methods [12–14] and graph convolutional neural network (GCN)-based methods [15–21]. Among these methods, GCN-based methods have gained much attention due to the natural graph structure of human skeletal data. Inspired by the spatial–temporal graph convolutional network (ST-GCN) introduced by Yan et al. [15], several GCN-based methods have been proposed to model the implicit relationships between skeletal data and the corresponding actions, thereby eliminating the need to design handcrafted part assignments or traversal rules.

In the field of action recognition, the information of each action is often concentrated on one or several vertices, and relies on the connections between them. Graph convolutions are commonly used to capture features from these joints and connections, and the adjacency matrix is used to determine the value of each connection, based on the physical connection conditions between the two vertices. However, in certain actions, two vertices may not be physically connected, but still be semantically correlated. For example, when someone

claps their hands, the vertices in the two hands are semantically correlated but physically unconnected. If the correlations between two vertices are simplified as physical connections, the flexibility and generalization abilities of recognition methods are inevitably limited. Thus, it is crucial to consider both physical and semantic connections when designing action recognition methods.

Unlike physical connections, that can be categorized as binary cases (i.e., connected and unconnected), the semantic correlation between two vertices is more complex. Firstly, the semantic correlation value should match the connection strength, making it continuous in the value space. Secondly, the semantic correlation is often independent of physical connections, and, thus their calculation cannot rely on the graph structure as priors. Attention mechanism can effectively highlight important features during the learning process. Therefore, we propose a novel approach to extensively incorporate semantic correlations in the graph convolutional calculation process, which introduces a novel two-stream graph attention convolutional network (2s-GATCN). The main contributions are as follows:

- We propose a graph attention convolutional network (GATCN) to adaptively learn the topology of the graph. By combining physical, semantic, and temporal features of the graph, our approach is able to learn and fuse features in a powerful and flexible manner.
- We present a novel approach for estimating semantic correlations by designing a graph attention block (GAB), which can highlight the most discriminative vertex connections relating to the corresponding actions. The GAB incorporates a data embedding method, to obtain a multi-channel semantic correlation strength tensor, and a CBAM-based attention module, to obtain the semantic correlation strength matrix, By considering the semantic connections between vertices, the action recognition accuracy is significantly improved.
- Extensive experiments on NTU-RGB+D 60 and Kinetics-Skeleton datasets demonstrate that the proposed network obtains superior performance compared to state-of-the-art methods.

The reminder of this paper is organized as follows: Section 2 discusses related work, Section 3 introduces the background, Section 4 presents the proposed 2s-GATCN framework, Section 5 presents experimental results, and Section 6 concludes the paper.

2. Related Work

2.1. Graph Covolutional Networks for Skeleton-Based Action Recognition

In recent years, GCN-based methods have been introduced to skeleton-based action recognition tasks by extending the convolutional operation from images to graphs, producing promising results in accurate action recognition from skeletal data.

The pioneering work was proposed by Yan et al. [15]. They proposed the spatial-temporal GCN (ST-GCN) to model the implicit relationships between action classes and skeletal data by proposing spatial graph convolution and temporal graph convolution. For spatial graph convolution, specific convolution kernels are designed, based on predefined principles, and the GCN calculation process is combined with an adjacency matrix that describes the graph's topological structure. Li et al. [17] proposed the actional–structural graph convolutional network (AS-GCN), which formulated generalized skeletal graphs by combining actional and structural links to capture action-specific latent dependencies. The actional links were obtained from an encoder–decoder structure, and the structural links were indicated by a high-order polynomial of the adjacency matrix. Shi et al. [16] introduced an adaptive learning method using a two-stream network (2s-AGCN) that utilized first-order joint information and second-order bone information to learn connection relationships. In 2s-AGCN, the adjacency matrix is extended with a learnable parameter matrix and a data-dependent connection matrix. In other work, Shi et al. [18] represented skeletal data as a directed acyclic graph (DAG), based on kinematic dependency, and designed a directed graph neural network (DGNN) to extract features from joints, bones and their relationships. Liu et al. [19] proposed a multi-scale aggregational scheme that

could remove redundant dependencies between near and far neighborhood joints, and proposed a unified spatial-temporal graph convolutional (G3D) operator that could directly extract spatial and temporal features from skeleton sequences. Peng et al. [20] proposed a neural searching-based approach that described implicit joint connections using a high-level representation of the skeleton graph and a dynamic graph modeling mechanism. To enrich the GCN search space, multiple dynamic graph substructures were provided, and higher-order connections with the Chebyshev polynomial approximation were applied. Chen et al. [21] proposed a channel-wise topology refinement graph convolutional network (CTR-GCN), which dynamically models channel-wise typologies in a refinement approach. The adjacency matrix was used as a shared topology for all channels, and the channel-specific correlations were used as a non-shared topology for each channel. Zhao et al. [22] introduced two progressive binary graph convolutional networks, in which the filters and activations were binarized to decrease the parameters, which could improve the training and inference speed. Zhang [23] proposed a spatial attentive and temporal dilated (SATD) method, in which the spatial attention pooling module (SAP) is proposed to identify important vertices and to remove unimportant vertices, and the temporal dilated graph convolution module is used to expand the receptive field. Yang [24] designed a hybrid network (HybridNet) , which integrated GCNs and CNNs to leverage their complementary effects. In HybridNet, a GCN-based feature extracting module (GFEM) and a CNN-based feature processing module (CFPM) were designed, respectively, and, then, a novel gluing unit was proposed to support the elegant integration of the GFEM and the CFEM.

2.2. Attention Mechanisms in Skeleton-Based Action Recognition

The attention mechanism has become a widely applied technique in various fields, including computer vision [25], video processing [26], and knowledge concept recommendation [27]. For skeleton-based action recognition, the attention module is an essential component that aims to identify and emphasize critical vertices, edges, or connections. Various attention models have been proposed to enhance recognition performance. Zhang et al. [28] introduced a regularized attention model that identified key vertices of each action by considering spatial diversity and local continuity. Song et al. [8] proposed an end-to-end spatial-temporal attention model, in which the spatial attention module aimed to assign different importance to each vertex, and the temporal attention module aimed to allocate different attention weights to each frame. Si et al. [29] proposed the AGC-LSTM network, in which an attention mechanism was used to highlight the features of key vertices, which could improve spatial-temporal expressions. Cho et al. [30] proposed three self-attention networks to effectively capture deep correlations from action sequences, which could address the issue of acquiring long-term information. Li et al. [31] incorporated a memory attention network into the 'RNN + CNN' network framework, which could effectively extracts the temporal features.

Currently, many attention modules are designed for non-GCN-based models, which limits their abilities to estimate the connection properties based on graph structures. Therefore, it is crucial to develop attention modules that can be used effectively with GCNs. One such module is the graph attention module (GAT), proposed by Velickovic et al [32]. The GAT module estimated the connection weights of different vertices, which could be used for inductive tasks. Yang et al. [33] proposed a pseudo graph convolutional network with temporal- and channel-wise attention, from which mixed temporal- and channel-wise attention was proposed to extract the different levels of importance of different frames and channels. Heidari et al. [34] proposed a temporal attention-augmented GCN by introducing a temporal attention module (TAM) to extract the most informative skeletons in an action sequence.

3. Background

3.1. Graph-Based Skeleton Sequence Representation

A skeleton sequence can be denoted as $X \in \mathbb{R}^{d \times T \times N}$, where d is the joint coordinate dimension, T is the frame number, and N is the number of joints in a skeletal frame. By utilizing the joint and joint connection information, there are two approaches that transform X into a graph, denoted as $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the vertices, and $\mathcal{E}$ represents the edges. For the joint graph, $\mathcal{V}$ represents a set of all the body joints, and $\mathcal{E}$ represents a set of all the bones calculated by the first-order spatial difference of the body joints. On the other hand, for the bone graph, the bones are set as the vertices, and the joints are set as the edges. The research task of this paper was to recognize specific human actions from the joint and bone graphs by applying the graph convolutions.

3.2. Spatial and Temporal Graph Convolutions

The spatial and temporal graph convolution operations were introduced by [15], which performed different convolutions on spatial and temporal dimensions based on the input feature with joint or bone information. In the spatial dimension, the graph convolution operation is expressed as:

$$f_{out} = \sum_{v_j^t \in \mathcal{B}_i} \frac{1}{Z_{ij}^t} f_{in}(v_j^t) \cdot \omega(l_i^t(v_j^t)), \tag{1}$$

where $f_{in} \in \mathbb{R}^{C_{in} \times T \times N}$ is the input feature map, and C_{in} is the input feature channel dimension. $f_{out} \in \mathbb{R}^{C_{out} \times T \times N}$ is the output feature map, and C_{out} is the output feature channel dimension. i is the target vertex index, j is the 1-distance neighbor vertex index of the target vertex i, and t is the frame index. $\mathcal{B}_i$ represents the sampling area of the graph convolution for v_i^t, which enumerates the vertices v_j^t. The value $\omega(\cdot)$ is the convolutional weight coefficient, $l_i^t(\cdot)$ is a mapping function that assigns convolutional weight coefficients to each involved vertex, and $Z_i^t(\cdot)$ is a normalization function that balances the contribution of the involved vertices.

It is difficult to directly utilize Equation (1) to extract spatial features from the graph. To better accommodate the graph topology, Equation (1) is typically transformed to Equation (2) according to [35].

$$f_{out} = \sum_{k}^{K_v} W_k f_{in} A_k, \tag{2}$$

where K_v is the spatial kernel size and set to 3, $W(\cdot)$ is the weight coefficient of a 1×1 convolutional layer, and $A \in \mathbb{R}^{N \times N}$ is an adjacency matrix that denotes the physical connections between the vertices. To perform the graph convolutions, f_{in} is first reshaped into a matrix with a size of $C_{in}T \times N$, and, after processing through the graph convolutional layers, the output feature is reshaped back to a tensor with size of $C_{out} \times T \times N$.

To perform graph convolutions along the temporal dimension, $K_t \times 1$ convolutions are applied, which is similar to the temporal convolution operations in video data processing. Here, K_t refers to the kernel size of the convolution operation along the temporal dimension.

4. Methodology

4.1. Overall Network

To fully utilize the graph feature, the proposed network adopts a 2-stream network framework that consists of joint and bone layers. The joint and bone layers are both structured as the proposed GATCN, as shown in Figure 1. Each layer produces a human action prediction score, and the two scores are summed to obtain the final prediction score.

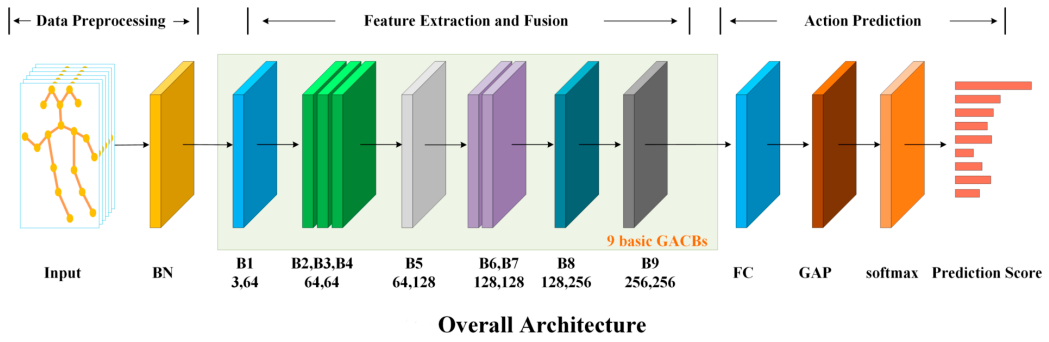

Figure 1. The Overall network of GATCN, which takes joints as example input.

The GATCN can be divided into two sub-blocks: the feature extraction and fusion block, and the action prediction block. The input to the network is the vertex set $\mathcal{V} \in \mathbb{R}^{d \times T \times N}$, which is first preprocessed using a batch normalization (BN) layer to normalize the data. Then, the normalized data $\mathcal{V}_n$ is fed into cascaded basic graph attention convolutional blocks (basic-GACBs) to extract discriminative graph features. Taking 3D joints as example input, the feature extraction and fusion process can be expressed as:

$$\begin{cases} f_1 = H_G(\mathcal{V}_n)_{C_{in}=3}^{C_{out}} \\ f_{i+1} = H_G(f_i)_{C_{in}}^{C_{out}}, \end{cases} \tag{3}$$

where f_i is the feature map extracted by the ith GACB, H_G is the GACB operation, and C_{in} and C_{out} are the input and output feature dimensions of the Basic-GACB,

Finally, the extracted feature is fed into the action prediction block to predict the action class, which consists of a fully connected (FC) layer, a global averaging pooling (GAP) layer and a softmax layer. The FC layer maps the features to the desired output size. The result is then fed to the GAP layer to reduce the spatial dimension of the features and compute the mean value of each feature map. Finally, the softmax layer is applied to obtain the predicted probability distribution over the action classes.

4.2. GACB

The primary block of the feature extraction and fusion block is Basic-GACB, as shown in Figure 2. Each GACB consists of two cascaded sub-blocks: a spatial feature extraction block and a temporal feature extraction block.

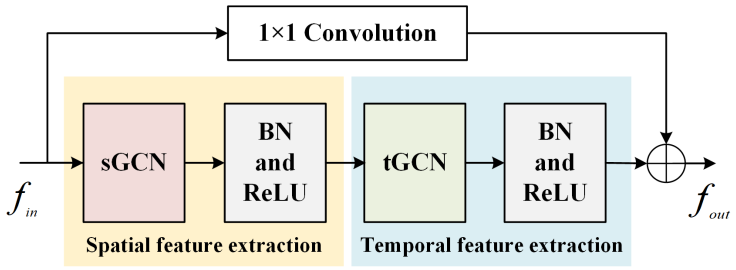

Figure 2. Basic-GACB.

The spatial feature extraction block of the Basic-GACB aims to extract spatial features by computing the adjacency matrix representing vertex connection strength. It consists of a spatial graph convolutional layer (sGCN), a BN layer, and a rectified linear unit (ReLU) layer. The calculation process of sGCN is illustrated in Figure 3. The adjacency matrix considers physical and semantic connection correlations and is divided into three parts: the physical connection matrix **A**, the semantic connection strength matrix **T**, and the learnable bias matrix **B**. Thus, the sGCN can be expressed as follows:

$$f_{out} = \sum_{k}^{K_v} \mathbf{W}_k f_{in}(\mathbf{A}_k + \mathbf{T}_k + \mathbf{B}_k), \tag{4}$$

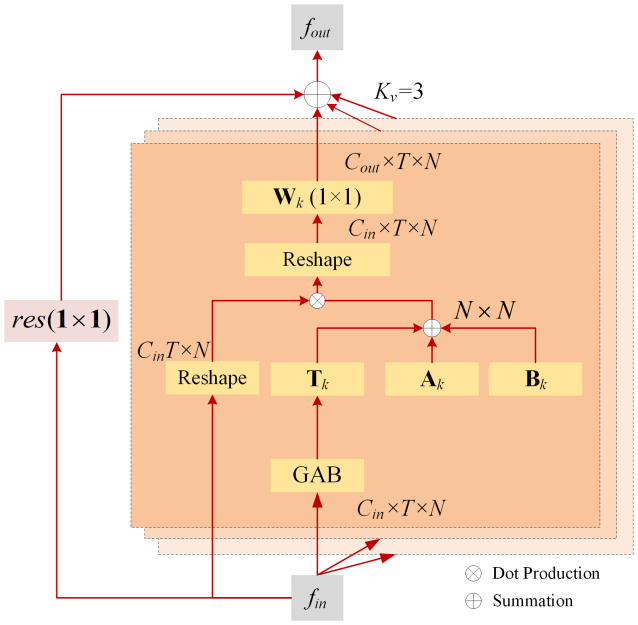

Figure 3. The calculation process of sGCN.

For the adjacency matrix, the physical connection matrix **A** is the same as that in Equation (2), which is predefined, based on the graph structure. It outputs 1 when the vertices are physically connected, and 0 otherwise. The semantic connection strength matrix **T** is obtained by applying the proposed graph attentional block (GAB) (as described in Section 4.3). This enables the model to identify the most discriminative vertex connections related to the corresponding actions. As the connection strength value of any two vertex is calculated and falls between 0 and 1, **T** can highlight the high-corrected semantic connections even when the two vertices are physically unconnected. The learnable bias matrix **B** is randomly initialized and learned during training. The value **B** also indicates the connection strength between the vertices, but its value is fixed for all input data. Therefore, it can be considered to demonstrate the overall connection property of the dataset.

The extracted spatial feature is then added to the input feature to formulate a global residual connection. Note that the input feature is processed with a 1×1 convolutional layer to make the channel dimension C_{out}.

The temporal feature extraction block is cascaded after the spatial graph convolutional block, and it consists of three layers: a temporal graph convolutional layer (tGCL), a BN layer and an Relu layer. This block takes the output of the spatial feature extraction block as input and applies temporal convolutions along the temporal frame dimension, resulting in a spatial-temporal fused feature map.

Finally, the spatial-temporal fused feature is added to the input feature to formulate a global residual connection. The input feature is also processed with a 1×1 convolutional layer to ensure the channel dimension is consistent with the spatial-temporal fused feature.

4.3. GAB

The GAB framework is designed to generate an attention matrix $\mathbf{T} \in R^{N\times N}$ that indicates the connection strengths between pairs of vertices. As shown in Figure 4, GAB consists of a feature embedding and fusion block, and a CBAM-based attention block.

For feature embedding and fusion, a 1×1 convolutional layer is adopted to embed the input feature into a tensor with a size of $C_e \times T \times N$, where C_e is the channel dimension of the embedded feature. Then, the matrix $\mathbf{M}_0$ with a size of $T \times C_e N$ is reshaped from the

embedded feature. Furthermore, f_{in} is fused with a 1×1 convolutional layer along the channel dimension to obtain an integrated feature $\mathbf{M}_1$ with a size of $T \times N$.

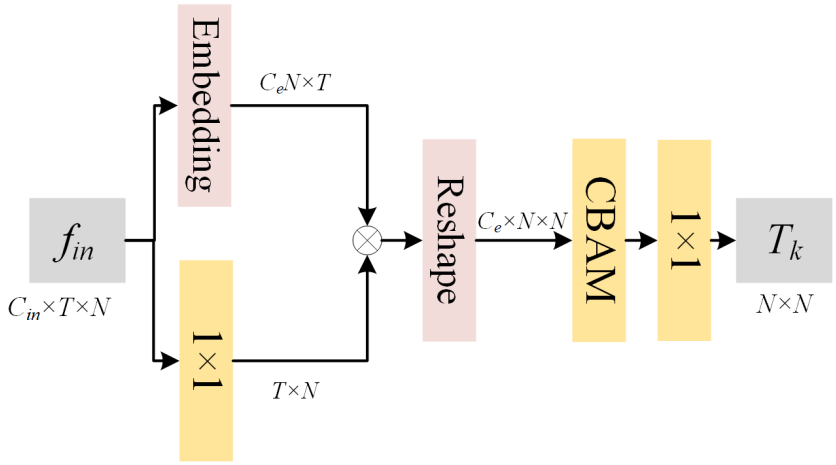

Figure 4. The calculation process of GAB.

A multi-channel connection strength tensor can be obtained by performing multiplication between M_0^T and M_1:

$$\mathbf{T}^m = R(\mathbf{M}_0^T \mathbf{M}_1), \tag{5}$$

where $\mathbf{T}^m = \{\mathbf{T}^1, \mathbf{T}^2, \cdots, \mathbf{T}^{C_e}\}$ is the multi-channel connection strength tensor with a size of $C_e \times N \times N$, m is the channel index and $R(\cdot)$ is the reshape operator. The element $\mathbf{T}^m(ij)$ indicates the connection strength between the vertex v_j and the vertex v_j in the m-th channel.

CBAM [36] is utilized to identify the most discriminative vertex connections in both the channel and spatial dimensions, as shown in Figure 5a. It consists of a channel attention block (CAB) and a spatial attention block (SAB), as illustrated in Figure 5b,c, respectively. The CAB aims to identify the most discriminative $\mathbf{T}^m$ along the channel dimension. First, the input feature is average-pooled and max-pooled in each $\mathbf{T}^m$. Then, the resulting features are respectively fed to a multilayer perceptron (MLP) layer to obtain two attention feature maps with a size of $C_e \times 1 \times 1$. Finally, the two attention feature maps are summed and passed through a sigmoid layer to obtain a channel attention matrix $\mathbf{A}_c$. The SAB aims to identify the most discriminative $\mathbf{T}(ij)$ along the spatial dimension. First, the input feature is average-pooled and max-pooled in the channel dimension. Then, the resulting features are concatenated and passed through a 7×7 convolutional layer to obtain a spatial attention matrix $\mathbf{A}_s$ with a size of $1 \times N \times N$. By multiplying $\mathbf{T}^m$ with $\mathbf{A}_c$ and $\mathbf{A}_s$ and fusing the resulting channels with a 1×1 convolutional layer, the semantic connection strength matrix $\mathbf{T}_{att}^m$ can be obtained as follows:

$$\mathbf{T}_{att}^m = H_{1\times1}(\mathbf{A}_s^m \mathbf{A}_c \mathbf{T}^m). \tag{6}$$

where $H_{1\times1}(\cdot)$ is the 1×1 convolution operation.

Finally, our GAB is designed as a multi-head mechanism to enhance the stability of the training process. The attention matrix from each head is summed to obtain the final attention matrix, which can be expressed as:

$$\mathbf{T} = \frac{1}{nheads} \sum_{i=1}^{nheads} \mathbf{T}_{att}^i. \tag{7}$$

where $\mathbf{T}$ is the final semantic connection strength matrix, $\mathbf{T}_{att}^i$denotes the attention matrix calculated with ith head, and $nhead$ is the number of heads.

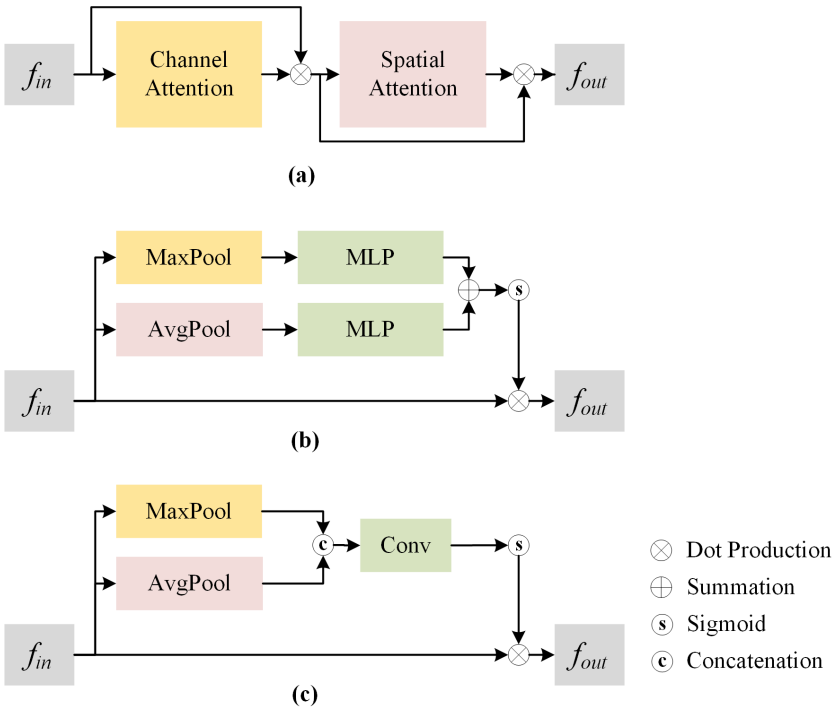

Figure 5. (**a**) The convolutional block attention module (CBAM). (**b**) The channel attention block. (**c**) The spatial attention block.

5. Experiments

In this section, we first introduce the datasets and implementation details used in our experiments. We then conduct an ablation study to evaluate the effectiveness of our proposed network. Finally, we compare our method with several state-of-the-art methods.

5.1. Datasets and Implementing Details

We used the NTU-RGB+D 60 and Kinetics-Skeleton datasets to train and test our network. The NTU-RGB+D 60 dataset contains 56,880 motion samples with 3D skeletal data and includes 60 action classes. It is divided into two benchmarks: Cross-view and Cross-subject. The Kinetics-Skeleton dataset is comprised of approximately 500,000 video clips covering 600 classes of human actions, with each class containing more than 600 video clips.

The network was implemented using the Pytorch framework, with a batch size of 20, and optimized by the stochastic gradient descent (SGD) optimizer, with an initial learning rate of 0.01. The GATCN was composed of 9 Basic-GACBs, each with an output feature dimension of 64, 64, 64, 64, 128, 128, 128, 256 and 256. To prevent over-fitting, L_2 regularization with a parameter of 0.0005 and a dropout rate of 0.6 were employed. The training process for the NTU-RGB+D 60 dataset was stopped after approximately 50 epochs, while for the Kinetics-Skeleton dataset, the training process was stopped after approximately 65 epochs.

5.2. Ablation Study

In this subsection, the Cross-view benchmark of the NTU-RGB+D 60 dataset was employed to identify the optimal hyperparameter within the network and to evaluate the effectiveness of the proposed semantic connection strength matrix T.

5.2.1. Optimal Attention Heads Number Determination

The proposed GAB utilizes multi-head GAB to enhance the stability of the learning process. In order to determine the optimal number of attention heads (*nheads*), we conducted experiments on the joint GATCN by varying *nheads* from 1 to 6. The corresponding results

are presented in Table 1, where the best and second best results are highlighted in **bold** and *underlined*. Based on the results, we observed that the recognition accuracy was highest when *nheads* was set to 3. Considering both recognition accuracy and computational efficiency, we recommend using $nhead = 3$ in GATCN.

Table 1. The validation accuracy comparisons of models using GAB with different numbers of heads.

nheads	**1**	**2**	**3**	**4**	**5**	**6**
accuracy(%)	94.43	94.45	**94.61**	94.53	94.55	*94.60*

5.2.2. Effectiveness Validation of the Matrix **T**

Two experiments were conducted to investigate the effectiveness of the proposed semantic connection strength matrix **T**, the results of which are listed in Table 2. We first deleted the matrix **T** to observe the performance. Based on the comparative results, the accuracy decreased by 1.1% in the joint stream network (Js-GATCN) and 1.2% in the bone stream network (Bs-GATCN). This is because the matrix **T** enables an effective description of connection characteristics between any two vertices. By highlighting the most discriminative vertex connections, the network is able to extract semantically high-correlated features more effectively.

Table 2. Effectiveness validation of the matrix **T**.

Method	**Accuracy(%)**	**Method**	**Accuracy(%)**
Js-GATCN $(\mathbf{A}+\mathbf{B})$	93.5	Bs-GATCN $(\mathbf{A}+\mathbf{B})$	93.3
Js-GATCN $(\mathbf{A}+\mathbf{B}+\mathbf{C})$	*93.8*	Bs-GATCN$(\mathbf{A}+\mathbf{B}+\mathbf{C})$	*93.5*
Js-GATCN $(\mathbf{A}+\mathbf{B}+\mathbf{T})$	**94.6**	Bs-GATCN$(\mathbf{A}+\mathbf{B}+\mathbf{T})$	**94.5**

Then, we replaced the proposed **T** with **C**, which was proposed in 2s-AGCN [16] and was calculated using the normalized embedded Gaussian function. Based on the results, we observed a decrease in accuracy by 0.8% in the joint stream network and by 1.0% in the bone stream network. The comparative result demonstrated that the proposed **T** is more effective in describing the connection strength between the vertices, as the introduction of the CBAM module.

5.3. Comparison With State-of-the-Art Methods

In this subsection, we compared the proposed method with several state-of-the-art skeleton-based action recognition methods on the NTU-RGB+D 60 and Kinetics-Skeleton datasets. Table 3 presents the comparison results on the NTU-RGB+D 60 dataset with the evaluation metric of top-1 classification accuracy. The results show that our method achieved the best score in both the Cross-subject and Cross-view benchmarks. Table 4 shows the comparison results on the Kinetics-Skeleton dataset, where we evaluated the model based on its top-1 and top-5 classification accuracy metrics. Our method achieved the best score in both metrics, demonstrating its effectiveness on a large dataset. Overall, the comparative results on these two datasets demonstrate the superiority of our method over the state-of-the-art approaches.

Table 3. Comparisons of the validation accuracy with state-of-the-art methods on NTU-RGB+D 60 dataset.

Methods	Year	Cross-Subject	Cross-View
HBRNN [10]	2015	50.1	82.8
ST-LSTM [37]	2016	69.2	77.7
Two-Stream 3DCNN [38]	2017	66.8	72.6
TCN [39]	2017	74.3	83.1
ST-GCN [15]	2018	81.5	88.3
AS-GCN [17]	2018	86.8	94.2
RA-GCN [40]	2019	85.9	93.5
2s-AGCN [16]	2019	88.5	95.1
AGC-LSTM [29]	2019	89.2	95.0
SGN [41]	2020	89.0	94.5
PL-GCN [42]	2020	89.2	90.5
SAGN [43]	2021	89.2	94.2
ED-GCN [44]	2022	88.7	95.2
GAT [45]	2022	89.0	95.2
Zhu [46]	2022	**89.6**	94.9
Js-GATCN	-	87.9	94.6
Bs-GATCN	-	87.4	94.5
2s-GATCN	-	**89.6**	**95.9**

Table 4. Comparisons of the validation accuracy with state-of-the-art methods on Kinetics-Skeleton dataset.

Method	Year	Top-1(%)	Top-5(%)
TCN [39]	2017	20.3	40.0
ST-GCN [15]	2018	30.7	52.8
AS-GCN [17]	2018	34.8	56.5
2s-AGCN [16]	2019	36.1	58.7
GAT [45]	2022	36.1	58.9
Zhu [46]	2022	34.0	57.5
Js-GATCN	-	34.1	57.2
Bs-GATCN	-	34.7	56.6
2s-GATCN	-	**36.7**	**59.8**

6. Conclusions

In this paper, we propose a novel two-stream graph attentional convolutional network (2s-GATCN) to improve the performance of skeleton-based human action recognition. The main contribution of our approach is the introduction of a graph attention block (GAB), which consists of an improved data embedding block and a CBAM-based attention block. The GAB generates a semantic connection strength matrix to identify the most discriminative vertex connections related to specific human actions, which allows for efficient extraction of features with high correlations in semantic space. We evaluated our model on the NTU-RGB+D 60 and Kinetics-Skeleton datasets and showed competitive performance compared to state-of-the-art methods. However, our method mainly focuses on optimizing the feature extraction and fusion process in the spatial domain. In the future, we plan to address the problem of dealing with semantic correlations among temporal frames.

Author Contributions: Conceptualization, S.-B.Z. and X.-Q.J.; Methodology, S.-B.Z. and X.-Q.J.; Software, S.-B.Z. and R.-R.C.; Validation, S.-B.Z., R.-R.C. and F.P.; Formal analysis, S.-B.Z.; Investigation, R.-R.C.; Writing—original draft, S.-B.Z.; Writing—review and editing, X.-Q.J. and F.P.; Supervision, X.-Q.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded partly by National Natural Science Foundation of China: 61803372, and partly by Natural Science Foundation of Shanghai: 20ZR140070.

Data Availability Statement: All data, models, or code supporting the results of this study are available from the authors upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A comprehensive survey of vision-based human action recognition methods. *Sensors* **2019**, *19*, 1005. [CrossRef] [PubMed]
2. Ushapreethi, P.; Jeyakumar, B.; BalaKrishnan, P. Action recongnition in video survillance using hipi and map reducing model. *Int. J. Mech. Eng. Technol.* **2017**, *8*, 368–375.
3. Ren, B.; Liu, M.; Ding, R.; Liu, H. A survey on 3d skeleton-based action recognition using learning method. *arXiv* **2020**, arXiv:2002.05907.
4. Ma, Z.; Liu, S. A review of 3D reconstruction techniques in civil engineering and their applications. *Adv. Eng. Inform.* **2018**, *37*, 163–174. [CrossRef]
5. Tian, F.; Gao, Y.; Fang, Z.; Fang, Y.; Gu, J.; Fujita, H.; Hwang, J.N. Depth estimation using a self-supervised network based on cross-layer feature fusion and the quadtree constraint. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1751–1766. [CrossRef]
6. Poppe, R. A survey on vision-based human action recognition. *Image Vis. Comput.* **2010**, *28*, 976–990. [CrossRef]
7. Jegham, I.; Khalifa, A.B.; Alouani, I.; Mahjoub, M.A. Vision-based human action recognition: An overview and real world challenges. *Forensic Sci. Int. Digit. Investig.* **2020**, *32*, 200901. [CrossRef]
8. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Voume 31.
9. Peng, W.; Shi, J.; Zhao, G. Spatial temporal graph deconvolutional network for skeleton-based human action recognition. *IEEE Signal Process. Lett.* **2021**, *28*, 244–248. [CrossRef]
10. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
11. Zhang, S.; Yang, Y.; Xiao, J.; Liu, X.; Yang, Y.; Xie, D.; Zhuang, Y. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks. *IEEE Trans. Multimed.* **2018**, *20*, 2330–2343. [CrossRef]
12. Du, Y.; Fu, Y.; Wang, L. Skeleton based action recognition with convolutional neural network. In Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 579–583.
13. Zheng, W.; Li, L.; Zhang, Z.; Huang, Y.; Wang, L. Relational network for skeleton-based action recognition. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 826–831.
14. Ding, W.; Ding, C.; Li, G.; Liu, K. Skeleton-Based Square Grid for Human Action Recognition With 3D Convolutional Neural Network. *IEEE Access* **2021**, *9*, 54078–54089. [CrossRef]
15. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
16. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 12026–12035.
17. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 3595–3603.
18. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with directed graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2021; pp. 7912–7921.
19. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 143–152.
20. Peng, W.; Hong, X.; Chen, H.; Zhao, G. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Voume 34, pp. 2669–2676.

21. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13359–13368.
22. Zhao, M.; Dai, S.; Zhu, Y.; Tang, H.; Xie, P.; Li, Y.; Liu, C.; Zhang, B. PB-GCN: Progressive binary graph convolutional networks for skeleton-based action recognition. *Neurocomputing* **2022**, *501*, 640–649. [CrossRef]
23. Zhang, J.; Ye, G.; Tu, Z.; Qin, Y.; Qin, Q.; Zhang, J.; Liu, J. A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. *CAAI Trans. Intell. Technol.* **2022**, *7*, 46–55. [CrossRef]
24. Yang, W.; Zhang, J.; Cai, J.; Xu, Z. HybridNet: Integrating GCN and CNN for skeleton-based action recognition. *Appl. Intell.* **2023**, *53*, 574–585. [CrossRef]
25. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
26. Liu, Y.; Zhang, X.; Huang, F.; Zhang, B.; Li, Z. Cross-attentional spatio-temporal semantic graph networks for video question answering. *IEEE Trans. Image Process.* **2022**, *31*, 1684–1696. [CrossRef]
27. Gong, J.; Wang, S.; Wang, J.; Feng, W.; Peng, H.; Tang, J.; Yu, P.S. Attentional graph convolutional networks for knowledge concept recommendation in moocs in a heterogeneous view. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 25–30 July 2020; pp. 79–88.
28. Zhang, G.; Zhang, X. Multi-heads attention graph convolutional networks for skeleton-based action recognition. In Proceedings of the 2019 IEEE Visual Communications and Image Processing (VCIP), Sydney, Australia, 1–4 December 2019; pp. 1–4.
29. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1227–1236.
30. Cho, S.; Maqbool, M.; Liu, F.; Foroosh, H. Self-attention network for skeleton-based human action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 635–644.
31. Li, C.; Xie, C.; Zhang, B.; Han, J.; Zhen, X.; Chen, J. Memory attention networks for skeleton-based action recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 4800–4814. [CrossRef]
32. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *Stat* **2017**, *1050*, 10–48550.
33. Yang, H.; Yan, D.; Zhang, L.; Sun, Y.; Li, D.; Maybank, S.J. Feedback graph convolutional network for skeleton-based action recognition. *IEEE Trans. Image Process.* **2021**, *31*, 164–175. [CrossRef]
34. Heidari, N.; Iosifidis, A. Temporal attention-augmented graph convolutional network for efficient skeleton-based human action recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), MiCo Milano Congress Center, Italy, 10–15 January 2021; pp. 7907–7914.
35. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
36. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
37. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part III 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 816–833.
38. Liu, H.; Tu, J.; Liu, M. Two-stream 3d convolutional neural network for skeleton-based action recognition. *arXiv* **2017**, arXiv:1705.08106.
39. Soo Kim, T.; Reiter, A. Interpretable 3d human action analysis with temporal convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 20–28.
40. Song, Y.F.; Zhang, Z.; Wang, L. Richly activated graph convolutional network for action recognition with incomplete skeletons. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1–5.
41. Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; Zheng, N. Semantics-guided neural networks for efficient skeleton-based human action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1112–1121.
42. Huang, L.; Huang, Y.; Ouyang, W.; Wang, L. Part-level graph convolutional network for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Voume 34, pp. 11045–11052.
43. Fu, Z.; Liu, F.; Zhang, J.; Wang, H.; Yang, C.; Xu, Q.; Qi, J.; Fu, X.; Zhou, A. SAGN: Semantic adaptive graph network for skeleton-based human action recognition. In Proceedings of the 2021 International Conference on Multimedia Retrieval, Taipei, Taiwan, 21–24 August 2021; pp. 110–117.
44. Alsarhan, T.; Ali, U.; Lu, H. Enhanced discriminative graph convolutional network with adaptive temporal modelling for skeleton-based action recognition. *Comput. Vis. Image Underst.* **2022**, *216*, 103348. [CrossRef]

45. Zhang, J.; Xie, W.; Wang, C.; Tu, R.; Tu, Z. Graph-aware transformer for skeleton-based action recognition. *Vis. Comput.* **2022**, 1–12. [CrossRef]
46. Zhu, Q.; Deng, H.; Wang, K. Skeleton Action Recognition Based on Temporal Gated Unit and Adaptive Graph Convolution. *Electronics* **2022**, *11*, 2973. [CrossRef]