An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition

Chenyang Si[1,3] Wentao Chen[1,4] Wei Wang[1,3*] Liang Wang[1,2,3] Tieniu Tan[1,2,3,4]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR)
[2]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[3]University of Chinese Academy of Sciences (UCAS)
[4]University of Science and Technology of China (USTC)
{chenyang.si, wentao.chen}@cripac.ia.ac.cn, {wangwei, wangliang, tnt}@nlpr.ia.ac.cn

Abstract

Skeleton-based action recognition is an important task that requires the adequate understanding of movement characteristics of a human action from the given skeleton sequence. Recent studies have shown that exploring spatial and temporal features of the skeleton sequence is vital for this task. Nevertheless, how to effectively extract discriminative spatial and temporal features is still a challenging problem. In this paper, we propose a novel Attention Enhanced Graph Convolutional LSTM Network (AGC-LSTM) for human action recognition from skeleton data. The proposed AGC-LSTM can not only capture discriminative features in spatial configuration and temporal dynamics but also explore the co-occurrence relationship between spatial and temporal domains. We also present a temporal hierarchical architecture to increases temporal receptive fields of the top AGC-LSTM layer, which boosts the ability to learn the high-level semantic representation and significantly reduces the computation cost. Furthermore, to select discriminative spatial information, the attention mechanism is employed to enhance information of key joints in each AGC-LSTM layer. Experimental results on two datasets are provided: NTU RGB+D dataset and Northwestern-UCLA dataset. The comparison results demonstrate the effectiveness of our approach and show that our approach outperforms the state-of-the-art methods on both datasets.

1. Introduction

In computer vision, human action recognition plays a fundamental and important role, with the purpose of predicting the action classes from videos. Additionally, it has been studied for decades and is still very popular due to its extensive potential applications, *e.g.*, video surveillance, human-computer interaction, sports analysis and so on [17, 33, 1].

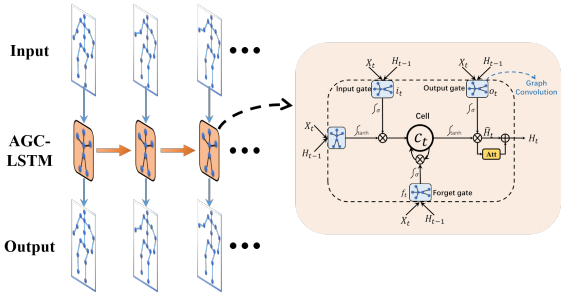


Figure 1. The structure of one AGC-LSTM layer. Different from traditional LSTM, the graph convolutional operator within AGC-LSTM causes the input, hidden state, and cell memory of AGC-LSTM to be graph-structured data.

Action recognition is a challenging task in the computer vision community. There are various attempts on human action recognition based on RGB video and 3D skeleton data. The RGB video based action recognition methods [23, 32, 25] mainly focus on modeling spatial and temporal representations from RGB frames and temporal optical flow. Despite RGB video based methods have achieved promising results, there still exist some limitations, *e.g.*, background clutter, illumination changes, appearance variation, and so on. 3D skeleton data represents the body structure with a set of 3D coordinate positions of key joints. And skeleton sequence does not contain color information and is not affected by the limitations of RGB video. Such robust representation allows to model more discriminative tempo-

*Corresponding Author: Wei Wang

ral characteristics about human actions. Moreover, Johansson *et al*. [8] have given an empirical and theoretical basis that key joints can provide highly effective information about human motion. Besides, the Microsoft Kinect [37] and advanced human pose estimation algorithms [2] make it easier to gain skeleton data.

For skeleton based action recognition, the existing methods explore different models to learn spatial and temporal features of skeleton sequences. Song *et al*. [24] employ a spatial-temporal attention model based on LSTM to select discriminative spatial and temporal features. The Convolutional Neural Networks (CNNs) are used to learn spatial-temporal features from skeletons in [3, 13, 9]. Yan *et al*. [35] propose a spatial-temporal graph convolutional network (ST-GCN) for action recognition. Compared with ST-GCN [35], Si *et al*. [21] propose to utilize the graph neural network and LSTM to represent spatial and temporal information, respectively. In short, all these methods are trying to design an effective model that can identify spatial and temporal features of skeleton sequence. Nevertheless, how to effectively extract discriminative spatial and temporal features is still a challenging problem.

Generally, there are three notable characteristics for human skeleton sequences: 1) There are strong correlations between each node and its adjacent nodes so that the skeleton frames contain abundant body structural information. 2) Temporal continuity exists not only in the same joints (*e.g*., hand, wrist and elbow), but also in the body structure. 3) There is a co-occurrence relationship between spatial and temporal domains. In this paper, we propose a novel and general framework called attention enhanced graph convolutional LSTM network (AGC-LSTM) for skeleton-based action recognition, which improves the skeleton representation by synchronously learning spatiotemporal characteristics mentioned above.

The architecture of the proposed AGC-LSTM network is shown in Fig.2. Firstly, the coordinate of each joint is transformed into a spatial feature with a linear layer. Then we concatenate spatial feature and feature difference between two consecutive frames to compose an augmented feature. In order to dispel scale variance between both features, a shared LSTM is adopted to process each joint sequence. Next, we apply three AGC-LSTM layers to model spatial-temporal features. As shown in Fig.1, due to the graph convolutional operator within AGC-LSTM, it can not only effectively capture discriminative features in spatial configuration and temporal dynamics but also explore the co-occurrence relationship between spatial and temporal domains. More specially, the attention mechanism is employed to enhance the features of key nodes at each time step, which can promote AGC-LSTM to learn more discriminative features. For example, the features of "elbow", "wrist" and "hand" are very important for action "hand-shaking" and should be enhanced in the process of identifying the behavior. Inspired by spatial pooling in CNNs, we present a temporal hierarchical architecture with temporal average pooling to increase temporal receptive fields of the top AGC-LSTM layers, which boosts the ability to learn high-level spatiotemporal semantic features and significantly reduces the computational cost. Finally, we use the global feature of all joints and the local feature of focused joints from the last AGC-LSTM layer to predict the class of human actions. Although the joint-based model achieves the state-of-the-art results, we also explore the performance of the proposed model on the part level. For the part-based model, the concatenation of joints of each part serves as a node to construct the graph. Furthermore, the two-stream model based on joint and part can lead to further performance improvement.

The main contributions of this work are summarized as follows:

- We propose a novel and general AGC-LSTM network for skeleton-based action recognition, which is the first attempt of graph convolutional LSTM for this task.
- The proposed AGC-LSTM is able to effectively capture discriminative spatiotemporal features. More specially, the attention mechanism is employed to enhance the features of key nodes, which assists in improving spatiotemporal expressions.
- A temporal hierarchical architecture is proposed to boost the ability to learn high-level spatiotemporal semantic features and significantly reduce the computational cost.
- The proposed model achieves the state-of-the-art results on both NTU RGB+D dataset and Northwestern-UCLA dataset. We perform extensive experiments to demonstrate the effectiveness of our model.

2. Related Work

Neural networks with graph Recently, graph-based models have attracted a lot of attention due to the effective representation for the graph structure data [34]. Existing graph models mainly fall into two architectures. One framework called graph neural network (GNN) is the combination of graph and recurrent neural network. Through multiple iterations of message passing and states updating of nodes, each node captures the semantic relation and structural information within its neighbor nodes. Qi *et al*. [18] apply GNN to address the task of detecting and recognizing human-object interactions in images and videos. Li *et al*. [14] exploit the GNNs to model dependencies between roles and predict a consistent structured output for situation recognition. The other framework is graph convolutional

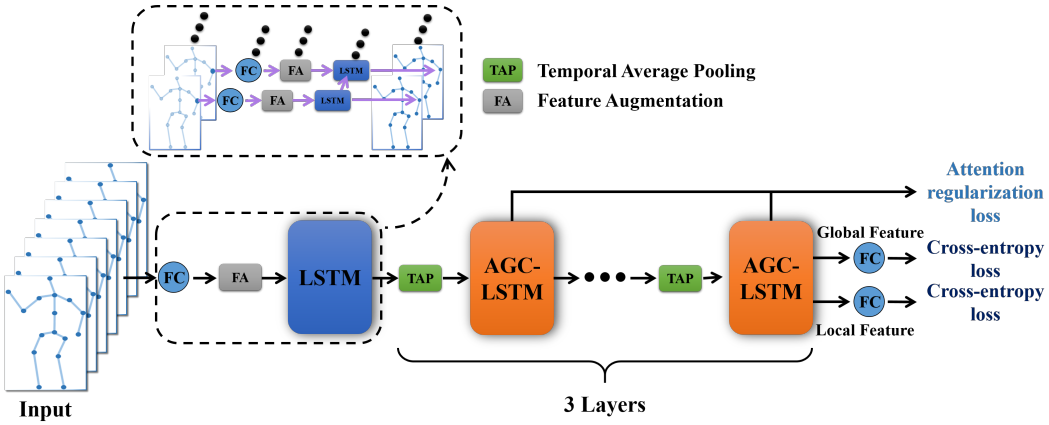


Figure 2. The architecture of the proposed attention enhanced graph convolutional LSTM network (AGC-LSTM). Feature augmentation (FA) computes feature differences with position features and concatenates both position features and feature differences. LSTM is used to dispel scale variance between feature differences and position features. Three AGC-LSTM layers can model discriminative spatial-temporal features. Temporal average pooling is the implementation of average pooling in the temporal domain. We use the global feature of all joints and the local feature of focused joints from the last AGC-LSTM layer to predict the class of human action.

network (GCN) that generalizes convolutional neural networks to graph. There are two types of GCNs: spectral GCNs and spatial GCNs. Spectral GCNs transform graph signals on graph spectral domains and then apply spectral filters on spectral domains. For example, the CNNs are utilized in the spectral domain relying on the graph Laplacian [5, 6]. Kipf *et al*. [11] introduce Spectral GCNs for semi-supervised classification on graph-structured data. For spatial GCNs, the convolution operation is applied to compute a new feature vector for each node using its neighborhood information. Simonovsky *et al*. [22] formulate a convolution-like operation on graph signals performed in the spatial domain and are the first to apply graph convolutions to point cloud classification. In order to capture the spatial-temporal features of graph sequences, a graph convolutional LSTM is firstly proposed in [19], which is an extension of GCNs to have the recurrent architecture. Inspired by [19], we exploit a novel AGC-LSTM network to learn inherent spatiotemporal representations from skeleton sequences.

Skeleton-based action recognition Human action recognition based on skeleton data has received a lot of attention, due to its effective representation of motion dynamics. Traditional skeleton-based action recognition methods mainly focus on designing hand-crafted features [26, 29, 7]. Vemulapalli *et al*. [27] represent each skeleton using the relative 3D rotations between various body parts. The relative 3D geometry between all pairs of body parts is applied to represent the 3D human skeleton in [26].

Recent works mainly learn human action representations with deep learning networks. Du *et al*. [4] divide human skeleton into five parts according to the human physical structure, and then separately feed them into a hierarchical recurrent neural network to recognize actions. A spatial-temporal attention network learns to selectively focus on discriminative spatial and temporal features in [24]. Zhang *et al*. [36] present a view adaptive model for skeleton sequence, which is capable of regulating the observation viewpoints to the suitable ones by itself. The works in [35, 13, 21] further show that learning discriminative spatial and temporal features is the key element for human action recognition. A hierarchical CNN model is presented in [13] to learn representations for joint co-occurrences and temporal evolutions. A spatial-temporal graph convolutional network (ST-GCN) is proposed for action recognition in [35]. Each spatial-temporal graph convolutional layer constructs spatial characteristics with a graph convolutional operator, and models temporal dynamic with a convolutional operator. Compared with ST-GCN [35], Si *et al*. [21] apply graph neural networks to capture spatial structural information and then use LSTM to model temporal dynamics. Despite the significant performance improvement in [21], it ignores the co-occurrence relationship between spatial and temporal features. In this paper, we propose a novel attention enhanced graph convolutional LSTM network that can not only effectively extract discriminative spatial and temporal features but also explore the co-occurrence relationship between spatial and temporal domains.

3. Model Architecture

In this section, we first briefly review the graph convolutional neural network and then introduce our attention enhanced graph convolutional LSTM. Finally, we illustrate the architecture of the proposed AGC-LSTM network.

3.1. Graph Convolutional Neural Network

Graph convolutional neural network (GCN) is a general and effective framework for learning representation of graph structured data. Various GCN variants have achieved the state-of-the-art results on many tasks. For skeleton-based action recognition, let $\mathcal{G}_t = \{\mathcal{V}_t, \mathcal{E}_t\}$ denotes a graph of human skeleton on a single frame at time t, where $\mathcal{V}_t$ is the set of N joint nodes and $\mathcal{E}_t$ is the set of skeleton edges. The neighbor set of a node v_{ti} is defined as $\mathcal{N}(v_{ti}) = \{v_{tj} | d(v_{ti}, v_{tj}) \leq D\}$, where $d(v_{ti}, v_{tj})$ is the minimum path length from v_{tj} to v_{ti}. A graph labeling function $\ell : \mathcal{V}_t \rightarrow \{1, 2, ..., K\}$ is designed to assign the labels $\{1, 2, ..., K\}$ to each graph node $v_{ti} \in \mathcal{V}_t$, which can partition the neighbor set $\mathcal{N}(v_{ti})$ of node v_{ti} into a fixed number of K subsets. The graph convolution is generally computed as:

$$\mathbf{Y}_{out}(v_{ti}) = \sum_{v_{tj} \in \mathcal{N}(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} \mathbf{X}(v_{tj}) \mathbf{W}(\ell(v_{tj})) \quad (1)$$

where $\mathbf{X}(v_{tj})$ is the feature of node v_{tj}. $\mathbf{W}(\cdot)$ is a weight function that allocates a weight indexed by the label $\ell(v_{tj})$ from K weights. $Z_{ti}(v_{tj})$ is the number of the corresponding subset, which normalizes feature representations. $\mathbf{Y}_{out}(v_{ti})$ denotes the output of graph convolution at node v_{ti}. More specifically, with the adjacency matrix, the Eqn. 1 can be represented as:

$$\mathbf{Y}_{out} = \sum_{k=1}^{K} \mathbf{\Lambda}_k^{-\frac{1}{2}} \mathbf{A}_k \mathbf{\Lambda}_k^{-\frac{1}{2}} \mathbf{X} \mathbf{W}_k \quad (2)$$

where $\mathbf{A}_k$ is the adjacency matrix in spatial configuration of the label $k \in \{1, 2, ..., K\}$. $\mathbf{\Lambda}_k^{ii} = \sum_j \mathbf{A}_k^{ij}$ is a degree matrix.

3.2. Attention Enhanced Graph Convolutional LSTM

For sequence modeling, a lot of studies have demonstrated that LSTM, as a variant of RNN, has an amazing ability to model long-term temporal dependencies. Various LSTM-based models are employed to learn temporal dynamics of skeleton sequences. However, due to the fully connected operator within LSTM, there is a limitation of ignoring spatial correlation for skeleton-based action recognition. Compared with LSTM, AGC-LSTM can not only capture discriminative features in spatial configuration and temporal dynamics, but also explore the co-occurrence relationship between spatial and temporal domains.

Like LSTM, AGC-LSTM also contains three gates: an input gate $\mathbf{i}_t$, a forget gate $\mathbf{f}_t$, an output gate $\mathbf{o}_t$. However, these gates are obtained with the graph convolution operator. The input $\mathbf{X}_t$, hidden state $\mathbf{H}_t$, and cell memory $\mathbf{C}_t$ of AGC-LSTM are graph-structured data. Fig.3 shows

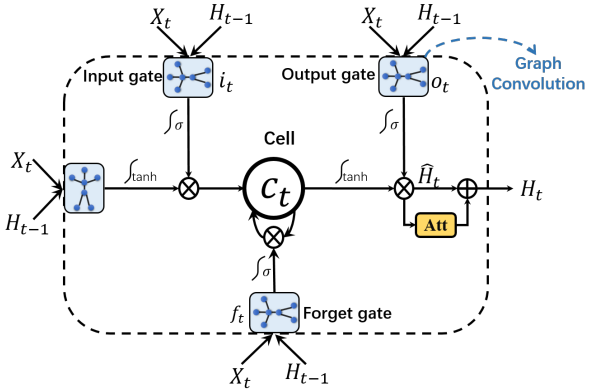


Figure 3. The structures of AGC-LSTM unit. Compared with LSTM, the inner operator of AGC-LSTM is graph convolutional calculation. To highlight more discriminative information, the attention mechanism is employed to enhance the features of key nodes.

the strcture of AGC-LSTM unit. Due to the graph convolutional operator within AGC-LSTM, the cell memory $\mathbf{C}_t$ and hidden state $\mathbf{H}_t$ not only exhibit temporal dynamics but also contain spatial structural information. The functions of AGC-LSTM unit are defined as follows:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}_{xi} *_{\mathcal{G}} \mathbf{X}_t + \mathbf{W}_{hi} *_{\mathcal{G}} \mathbf{H}_{t-1} + b_i) \\
\mathbf{f}_t &= \sigma(\mathbf{W}_{xf} *_{\mathcal{G}} \mathbf{X}_t + \mathbf{W}_{hf} *_{\mathcal{G}} \mathbf{H}_{t-1} + b_f) \\
\mathbf{o}_t &= \sigma(\mathbf{W}_{xo} *_{\mathcal{G}} \mathbf{X}_t + \mathbf{W}_{ho} *_{\mathcal{G}} \mathbf{H}_{t-1} + b_o) \\
\mathbf{u}_t &= tanh(\mathbf{W}_{xc} *_{\mathcal{G}} \mathbf{X}_t + \mathbf{W}_{hc} *_{\mathcal{G}} \mathbf{H}_{t-1} + b_c) \\
\mathbf{C}_t &= \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \mathbf{u}_t \\
\widehat{\mathbf{H}}_t &= \mathbf{o}_t \odot tanh(\mathbf{C}_t) \\
\mathbf{H}_t &= f_{att}\left(\widehat{\mathbf{H}}_t\right) + \widehat{\mathbf{H}}_t
\end{aligned} \quad (3)$$

where $*_{\mathcal{G}}$ denotes the graph convolution operator and $\odot$ denotes the Hadamard product. $\sigma(\cdot)$ is the sigmoid activation function. $\mathbf{u}_t$ is the modulated input. $\widehat{\mathbf{H}}_t$ is an intermediate hidden state. We use $\mathbf{W}_{xi} *_{\mathcal{G}} \mathbf{X}_t$ to mean a graph convolution of $\mathbf{X}_t$ with $\mathbf{W}_{xi}$, which can be written as Eqn.1. $f_{att}(\cdot)$ is an attention network that can select discriminative information of key nodes. The sum of $f_{att}\left(\widehat{\mathbf{H}}_t\right)$ and $\widehat{\mathbf{H}}_t$ as the output aims to strengthen information of key nodes without weakening information of non-focused nodes, which can maintain the integrity of spatial information.

The attention network is employed to adaptively focus on key joints with a soft attention mechanism that can automatically measure the importance of joints. The illustration of the spatial attention network is shown in Fig.4. The intermediate hidden state $\widehat{\mathbf{H}}_t$ of AGC-LSTM contains rich spatial structural information and temporal dynamics that are beneficial in guiding the selection of key joints. So we first

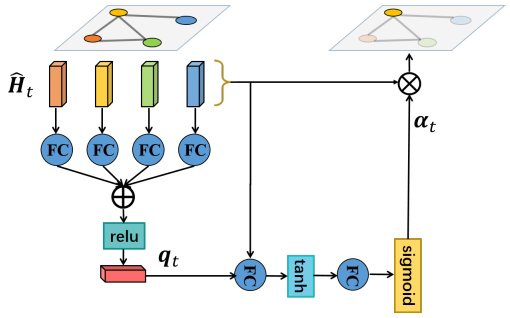


Figure 4. Illustration of the spatial attention network.

aggregate the information of all nodes as a query feature:

$$\mathbf{q}_t = ReLU\left(\sum_{i=1}^{N} \mathbf{W}\widehat{\mathbf{H}}_{ti}\right) \tag{4}$$

where $\mathbf{W}$ is the learnable parameter matrix. Then the attention scores of all nodes can be calculated as:

$$\boldsymbol{\alpha}_t = Sigmoid\left(\mathbf{U}_s tanh\left(\mathbf{W}_h\widehat{\mathbf{H}}_t + \mathbf{W}_q\mathbf{q}_t + \mathbf{b}_s\right) + \mathbf{b}_u\right) \tag{5}$$

where $\boldsymbol{\alpha}_t = (\alpha_{t1}, \alpha_{t2}, ..., \alpha_{tN})$, and $\mathbf{U}_s, \mathbf{W}_h, \mathbf{W}_q$ are the learnable parameter matrixes. $\mathbf{b}_s, \mathbf{b}_u$ are the bias. We use the non-linear function of *Sigmoid* due to the possibility of existing multiple key joints. The hidden state $\mathbf{H}_{ti}$ of node v_{ti} can also be represented as $(1+\alpha_{ti}) \cdot \widehat{\mathbf{H}}_{ti}$. The attention enhanced hidden state $\mathbf{H}_t$ will be fed into the next AGC-LSTM layer. Note that, at the last AGC-LSTM layer, the aggregation of all node features will serve as a global feature $\mathbf{F}_t^g$, and the weighted sum of focused nodes will serve as a local feature $\mathbf{F}_t^l$:

$$\mathbf{F}_t^g = \sum_{i=1}^{N} \mathbf{H}_{ti} \tag{6}$$

$$\mathbf{F}_t^l = \sum_{i=1}^{N} \alpha_{ti} \cdot \widehat{\mathbf{H}}_{ti} \tag{7}$$

The global feature $\mathbf{F}_t^g$ and local feature $\mathbf{F}_t^l$ are used to predict the class of human action.

3.3. AGC-LSTM Network

We propose an end-to-end attention enhanced graph convolutional LSTM network (AGC-LSTM) for skeleton-based human action recognition. The overall pipeline of our model is shown in Fig.2. In the following, we discuss the rationale behind the proposed framework in detail.

Joints Feature Representation. For the skeleton sequence, we first map the 3D coordinate of each joint into a high-dimensional feature space using a linear layer and an LSTM layer. The first linear layer encodes the coordinates of joints into a 256-dim vector as position features $\mathbf{P}_t \in \mathbb{R}^{N\times 256}$, and $\mathbf{P}_{ti} \in \mathbb{R}^{1\times 256}$ denotes the position representation of joint i. Due to only containing position information, the position feature $\mathbf{P}_{ti}$ is beneficial for learning spatially structured characteristic in the graph model. Frame difference features $\mathbf{V}_{ti}$ between two consecutive frames can facilitate the acquisition of dynamic information for AGC-LSTM. In order to take into account both advantages, the concatenation of both features serve as an augmented feature to enrich feature information. However, the concatenation of position feature $\mathbf{P}_{ti}$ and frame difference feature $\mathbf{V}_{ti}$ exists the scale variance of the features vectors. Therefore, we adopt an LSTM layer to dispel scale variance between both features:

$$\begin{aligned}\mathbf{E}_{ti} &= f_{lstm}\left(concat\left(\mathbf{P}_{ti}, \mathbf{V}_{ti}\right)\right) \\ &= f_{lstm}\left(concat\left(\mathbf{P}_{ti}, \left(\mathbf{P}_{ti} - \mathbf{P}_{(t-1)i}\right)\right)\right)\end{aligned} \tag{8}$$

where $\mathbf{E}_{ti}$ is the augmented feature of joint i at time t. Note that the linear layer and LSTM are shared among different joints.

Temporal Hierarchical Architecture. After the LSTM layer, the sequence $\{\mathbf{E}_1, \mathbf{E}_2, ..., \mathbf{E}_T\}$ of augmented features will be fed into the following GC-LSTM layers as the node features, where $\mathbf{E}_t \in \mathbb{R}^{N\times d_e}$. The proposed model stacks three AGC-LSTM layers to learn the spatial configuration and temporal dynamics. Inspired by spatial pooling in CNNs, we present a temporal hierarchical architecture of AGC-LSTM with average pooling in temporal domain to increase the temporal receptive field of the top AGC-LSTM layers. Through the temporal hierarchical architecture, the temporal receptive field of each time input at the top AGC-LSTM layer becomes a short-term clip from a frame, which can be more sensitive to the perception of the temporal dynamics. In addition, it can significantly reduce computational cost on the premise of improving performance.

Learning AGC-LSTM. Finally, the global feature $\mathbf{F}_t^g$ and local feature $\mathbf{F}_t^l$ of each time step are transformed into the scores $\mathbf{o}_t^g$ and $\mathbf{o}_t^l$ for C classes, where $\mathbf{o}_t = (o_{t1}, o_{t2}, ..., o_{tC})$. And the predicted probability being the i^{th} class is then obtained as:

$$\hat{y}_{ti} = \frac{e^{o_{ti}}}{\sum_{j=1}^{C} e^{o_{tj}}}, i = 1, ..., C \tag{9}$$

During training, considering that the hidden state of each time step on the top AGC-LSTM contains a short-term dynamics, we supervise our model with the following loss:

$$\begin{aligned}\mathcal{L} = &-\sum_{t=1}^{T_3}\sum_{i=1}^{C} y_i log\hat{y}_{ti}^g - \sum_{t=1}^{T_3}\sum_{i=1}^{C} y_i log\hat{y}_{ti}^l \\ &+ \lambda\sum_{j=1}^{3}\sum_{n=1}^{N}\left(1 - \frac{\sum_{t=1}^{T_j}\alpha_{tnj}}{T_j}\right)^2 + \beta\sum_{j=1}^{3}\frac{1}{T_j}\sum_{t=1}^{T_j}\left(\sum_{n=1}^{N}\alpha_{tnj}\right)^2\end{aligned} \tag{10}$$

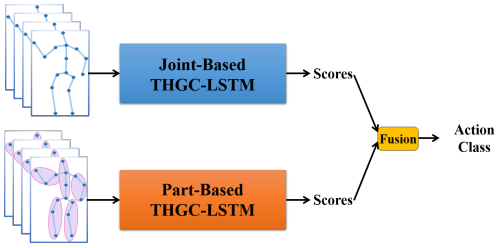


Figure 5. Illustration of the hybrid model based on joins and parts.

where $\boldsymbol{y} = (y_1, ..., y_C)$ is the groundtruth label. T_j denotes the number of time step on j^{th} AGC-LSTM layer. The third term aims to pay equal attention to different joints. The last term is to limit the number of interested nodes. λ and β are weight decaying coefficients. Note that only the sum probability of $\hat{\boldsymbol{y}}^g_{T_3}$ and $\hat{\boldsymbol{y}}^l_{T_3}$ at the last time step is used to predict the class of the human action.

Although the joint-based AGC-LSTM network has achieved the state-of-the-art results, we also explore the performance of the proposed model on the part level. According to human physical structure, the body can be divided into several parts. Similar to joint-based AGC-LSTM network, we first capture part features with a linear layer and a shared LSTM layer. Then the part features as node representations are fed into three AGC-LSTM layers to model spatial-temporal characteristics. The results illustrate that our model can also achieve superior performance on the part level. Furthermore, the hybrid model (shown in Fig.5) based on joints and parts can lead to further performance improvement.

4. Experiments

We have evaluated our proposed model on two datasets: NUT RGB+D dataset [20] and Northwestern-UCLA dataset [31]. The analysis of experimental results confirms the effectiveness of our model for skeleton-based action recognition.

4.1. Datasets

NTU RGB+D dataset [20]. This dataset contains 60 different human action classes that are divided into three major groups: daily actions, mutual actions, and health-related actions. There are 56,880 action samples in total which are performed by 40 distinct subjects. Each action sample contains RGB video, depth map sequence, 3D skeleton data, and infrared video captured by three Microsoft Kinect v2 cameras concurrently. The 3D skeleton data that we focus on consists of 3D positions of 25 body joints per frame. There are two evaluation protocols for this dataset: Cross-Subject (CS) and Cross-View (CV) [20]. Under the Cross-Subject protocol, actions performed by 20 subjects constitute the training set and the rest of actions performed by the other 20 subjects are used for testing. For Cross-View evaluation, samples captured by the first two cameras are used for training and the rest are for testing.

Northwestern-UCLA dataset [31]. This dataset contains 1494 video clips covering 10 categories. It is captured by three Kinect cameras simultaneously from a variety of viewpoints. Each action sample contains RGBD and human skeleton data performed by 10 different subjects. The evaluation protocol is the same as in [31]. Samples from the first two cameras constitute the training set and samples from the other camera constitute the testing dataset.

4.2. Implementation Details

In our experiments, we sample a fixed length T from each skeleton sequence as the input. We set the length $T = 100$ and 50 for NTU dataset and Northwestern-UCLA dataset, respectively. In the proposed AGC-LSTM, the neighbor set of each node contains only nodes directly connected with itself, so $D = 1$. In order to compare fairly with ST-GCN [35], the graph labeling function in AGC-LSTM will partition the neighbor set into $K = 3$ subsets according to [35]: the root node itself, centripetal group, and centrifugal group. The channels of three AGC-LSTM layers are set to 512. During training, we use the Adam optimizer [10] to optimize the network. Dropout with a probability of 0.5 is adopted to avoid over-fitting on these two datasets. We set λ and β to 0.01 and 0.001, respectively. The initial learning rate is set to 0.0005 and reduced by multiplying it by 0.1 every 20 epochs. The batch sizes for the NTU dataset and Northwestern-UCLA dataset are 64 and 30, respectively.

4.3. Results and Comparisons

In this section, we compare our proposed attention enhanced graph convolutional LSTM network (AGC-LSTM) with several state-of-the-art methods on the used two datasets.

4.3.1 NTU RGB+D Dataset

From Table 1, we can see that our proposed method achieves the best performance of 95.0% and 89.2% in terms of two protocols on the NTU dataset. To demonstrate the effectiveness of our method, we choose the following related methods to compare and analyze the results:

AGC-LSTM vs HCN. HCN [13] employs the CNN model for learning global co-occurrences from skeleton data. It treats each joint of a skeleton as a channel, then uses the convolution layer to learn the glob co-occurrence features from all joints. We can see that our performances significantly outperform the HCN [13] by about 3.9% and 2.7% for cross-view evaluation and cross-subject evaluation, respectively.

Methods	Year	CV	CS
HBRNN-L [4]	2015	64.0	59.1
Part-aware LSTM [20]	2016	70.3	62.9
Trust Gate ST-LSTM [15]	2016	77.7	69.2
Two-stream RNN [28]	2017	79.5	71.3
STA-LSTM [24]	2017	81.2	73.4
Ensemble TS-LSTM [12]	2017	81.3	74.6
Visualization CNN [16]	2017	82.6	76.0
VA-LSTM [36]	2017	87.6	79.4
ST-GCN [35]	2018	88.3	81.5
SR-TSL [21]	2018	92.4	84.8
HCN [13]	2018	91.1	86.5
AGC-LSTM (Joint)	-	93.5	87.5
AGC-LSTM (Part)	-	93.8	87.5
AGC-LSTM (Joint&Part)	-	**95.0**	**89.2**

Table 1. Comparison with the state-of-the-art methods on the NTU RGB+D dataset for Cross-View (CS) and Cross-Subject (CV) evaluation in accuracy.

AGC-LSTM vs ST-GCN. In order to compare fairly with [35], we use the same GCN operator in the proposed AGC-LSTM layer as in ST-GCN. For ST-GCN [35], it applies GCN to model spatial configuration of the joints, then uses the convolutional operator to learn temporal dynamics in each layer. On the joint-level evaluation, the results of AGC-LSTM are 93.5% and 87.5% that outperform 5.2% and 6.0% than ST-GCN. The comparison results prove that the AGC-LSTM is optimal for skeleton-based action recognition than ST-GCN.

Co-occurrence relationship between spatial and temporal domains. Although Si *et al*.[21] propose a spatial reasoning and temporal stack learning network with graph neural network (GNN) and LSTM, they ignore the co-occurrence relationship between spatial and temporal domains. Due to the ability to explore the co-occurrence relationship between spatial and temporal domains, Our AGC-LSTM outperforms [21] by 2.6% and 4.4%.

The performances on joint level and part level. Recent methods can be grouped into two categories: joint-based [35, 36, 12, 28, 13] and part-based methods [21, 28, 4]. Our method achieves the state-of-the-art results on joint-level and part-level, which illustrates the better generalization of our model for joint-level and part-level inputs.

4.3.2 Northwestern-UCLA Dataset

As shown in Table 2, the proposed AGC-LSTM again achieves the best accuracy of 93.3% on the Northwestern-UCLA dataset. The previous state-of-the-art model [12] employs multiple Temporal Sliding LSTM (TS-LSTM) to extract short-term, medium-term and long-term temporal dynamics respectively, which has similar functionality to our temporal hierarchical architecture. However, our model outperforms TS-LSTM [12] by 4.1%. Compared with the CNN-based method [16], our method also obtains much better performance.

Methods	Year	Accuracy (%)
Lie group [26]	2014	74.2
Actionlet ensemble [30]	2014	76.0
HBRNN-L [4]	2015	78.5
Visualization CNN [16]	2017	86.1
Ensemble TS-LSTM [12]	2017	89.2
AGC-LSTM (Joint)	-	92.2
AGC-LSTM (Part)	-	90.1
AGC-LSTM (Joint&Part)	-	**93.3**

Table 2. Comparison with the state-of-the-art methods on the Northwestern-UCLA dataset in accuracy.

4.4. Model Analysis

To understand the properties of our AGC-LSTM network, we analyze the effectiveness of several key components on both NTU RGB+D dataset and Northwestern-UCLA dataset, *i.e*. temporal hierarchical architecture, AGC-LSTM, attention enhanced mechanism in AGC-LSTM and the two-streams network. Finally, we analyze several failure cases to discuss the existing problems for skeleton-based action recognition.

4.4.1 Architecture Analysis

Tables 3 and 4 show experimental results of several baselines on the NTU RGB+D dataset and Northwestern-UCLA dataset, respectively. HT denotes temporal hierarchical architecture. Compared with LSTM and GC-LSTM, LSTM+HT and GC-LSTM+TH can increase the temporal receptive fields of each time step on the top layer. The improved performances prove that the temporal hierarchical architecture can boost the representation of temporal dy-

Methods		CV	CS
Joint	LSTM	89.4	80.3
	GC-LSTM	92.4	85.6
	LSTM+TH	90.4	81.4
	GC-LSTM+TH	92.9	86.3
	AGC-LSTM+TH (AGC-LSTM)	93.5	87.5
Part	AGC-LSTM+TH (AGC-LSTM)	93.8	87.5
AGC-LSTM (Joint&Part)		**95.0**	**89.2**

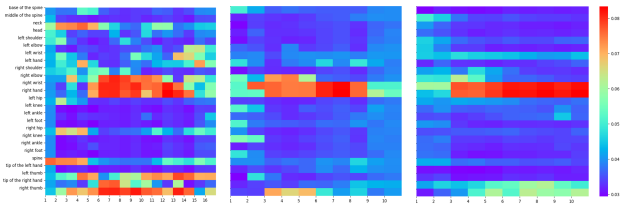
Table 3. The comparison results between several baselines and our AGC-LSTM on the NTU RGB+D dataset.

Methods		Accuracy (%)
Joint	LSTM	70.0
	GC-LSTM	87.5
	LSTM+TH	78.5
	GC-LSTM+TH	89.4
	AGC-LSTM+TH (AGC-LSTM)	92.2
Part	AGC-LSTM+TH (AGC-LSTM)	90.1
AGC-LSTM (Joint&Part)		**93.3**

Table 4. The comparison results between several baselines and our AGC-LSTM on the Northwestern-UCLA dataset.

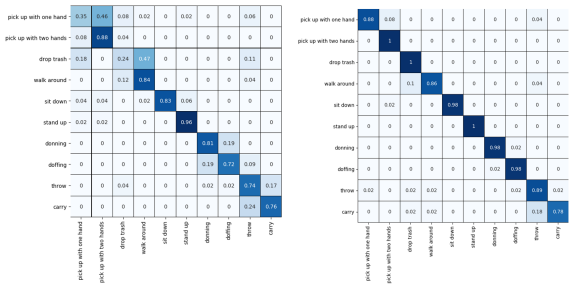
namics. Replacing LSTM with GC-LSTM, GC-LSTM+HT increases the accuracies to 2.5%, 4.9% on the NTU dataset and 10.9% on the Northwestern-UCLA dataset, respectively. Substantial performance improvements verify the effectiveness of GC-LSTM, which can capture more discriminative spatial-temporal features from skeleton data. Compared with GC-LSTM, AGC-LSTM can employ the spatial attention mechanism to select spatial information of key joints, which can promote the ability of feature representation. In addition, the fusion of part-based and joint-based AGC-LSTM can further improve the performance.

We also visualize the attention weights of three AGC-LSTM layers in Fig.6. For the "handshaking" action, the results show our method can gradually enhance the attention of "right elbow", "right wrist", and "right hand". Meanwhile, "tip of the right hand" and "right thumb" have some degree of attention. Furthermore, we analyze the experimental results with a confusion matrix on the Northwestern-UCLA dataset. As show in Fig.7(a), it is very confusing for LSTM to recognize similar actions. For example, the actions "pick up with one hand" and "pick up with two hands" have very similar skeleton sequences. Nevertheless, we can see that the proposed AGC-LSTM can significantly improve the ability to classify these similar actions (shown in Fig.7(b)). The above results illustrate that the proposed AGC-LSTM is an effective method for skeleton-based action recognition.



(a) (b) (c)

Figure 6. Visualizations of the attention weights of three AGC-LSTM layers on one actor of the action "handshaking". Vertical axis denotes the joints. Horizontal axis denotes the frames. (a), (b), (c) are the attention results of the first, second and third AGC-LSTM layer, respectively.



(a) LSTM (b) AGC-LSTM

Figure 7. Confusion matrix comparison on the Northwestern-UCLA dataset. (a) LSTM. (b) AGC-LSTM.

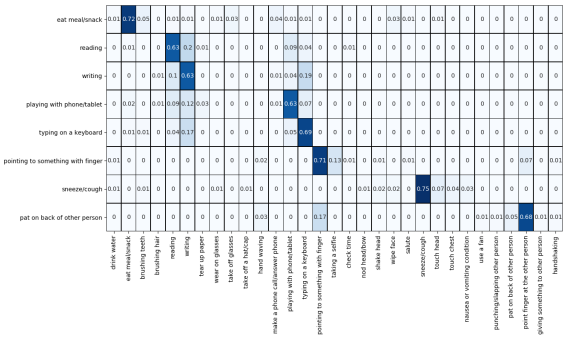


Figure 8. Confusion matrix comparison on the NTU dataset. It shows the part of confusion matrix comparison of the actions ("eat meal/snack", "reading", "writing", "playing with phone/tablet", "typing on a keyboard", "pointing to something with finger", "sneeze/cough", "pat on back of other person") with accuracies less than 80% on NTU dataset.

4.4.2 Failure Case

Finally, we analyze misclassification results with a confusion matrix on the NTU dataset. Fig.8 shows the part confusion matrix comparison of the actions ("eat meal/snack", "reading", "writing", "playing with phone/tablet", "typing on a keyboard", "pointing to something with finger", "sneeze/cough", "pat on back of other person") with accuracies less than 80% for the cross-subject setting on the NTU dataset. We can see that misclassified actions are mainly very similar movements. For example, 20% samples of "reading" are misclassified as "writing", and there are 19% sequences of "writing" misclassified as "typing on as keyboard". For the NTU dataset, only two joints are marked on fingers ("tip of the hand" and "thumb"), so that it is very challenging to capture such subtle movements of the hands.

5. Conclusion and Future Work

In this paper, we propose an attention enhanced graph convolutional LSTM network (AGC-LSTM) for skeleton-

based action recognition, which is the first attempt of graph convolutional LSTM for this task. The proposed AGC-LSTM can not only capture discriminative features in spatial configuration and temporal dynamics, but also explore the co-occurrence relationship between spatial and temporal domains. Furthermore, the attention network is employed to enhance information of key joints in each AGC-LSTM layer. In addition, we also propose a temporal hierarchical architecture to capture high-level spatiotemporal semantic features. On two challenging benchmarks, the proposed AGC-LSTM achieves the state-of-the-art results. Learning the pose-object relation could help overcome the limitations mentioned in the failure case. In the future, we will try the combination of skeleton sequence and object appearance to promote the performance of human action recognition.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 2011.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[3] Y. Du, Y. Fu, , and L. Wang. Skeleton based action recognition with convolutional neural network. In *ACPR*, 2015.

[4] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015.

[5] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*. 2015.

[6] M. Henaff, J. Bruna, and Y. LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

[7] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *IJCAI*, 2013.

[8] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 1973.

[9] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, 2017.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[11] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[12] I. Lee, D. Kim, S. Kang, and S. Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *ICCV*, 2017.

[13] C. Li, Q. Zhong, D. Xie, and S. Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *IJCAI*, 2018.

[14] R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, and S. Fidler. Situation recognition with graph neural networks. In *ICCV*, 2017.

[15] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, 2016.

[16] M. Liu, H. Liu, and C. Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 2017.

[17] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 2010.

[18] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018.

[19] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson. Structured sequence modeling with graph convolutional recurrent networks. *arXiv preprint arXiv:1612.07659*, 2016.

[20] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016.

[21] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In *ECCV*, 2018.

[22] M. Simonovsky and N. Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *CVPR*, 2017.

[23] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[24] 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*, 2017.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[26] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, 2014.

[27] R. Vemulapalli and R. Chellappa. Rolling rotations for recognizing human actions from 3d skeletal data. In *CVPR*, 2016.

[28] H. Wang and L. Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *CVPR*, 2017.

[29] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[30] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3d human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[31] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-view action modeling, learning, and recognition. In *CVPR*, 2014.

[32] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[33] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 2011.

[34] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *arXiv preprint arXiv:1810.00826*, 2018.

[35] S. Yan, Y. Xiong, D. Lin, and xiaoou Tang. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[36] 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 *ICCV*, 2017.

[37] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE Multimedia*, 2012.