# GRADIENT-WEIGHTED CLASS ACTIVATION MAPPING FOR SPATIO TEMPORAL GRAPH CONVOLUTIONAL NETWORK

*Pratyusha Das, Antonio Ortega*

Dept. of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA

## ABSTRACT

Spatio-temporal graph convolutional networks (STGCN) have become popular recently because they can handle structured data with dynamic temporal variations. However, the lack of interpretability limits the potential application of STGCNs. Gradient-based class activation maps (Grad-CAM) are a popular technique to interpret convolutional neural networks for grid structured data such as images. In this paper, we design an extension of Grad-CAMs for spatio temporal graph convolution (STG-Grad-CAM) to improve the interpretability of STGCNs. As a proof of concept we provide results for a skeleton-based activity recognition task. We show which body joints are responsible for a particular task and how their temporal dynamics contribute to the classification output. We present a brief study of the interpretability of a recognition task by changing the model depth and the training and testing protocol. To find the efficacy of STG-Grad-CAM, we compute faithfulness of STG-Grad-CAM to the model measured by the impact of occlusions to the graph nodes. For explainability of STGCN, we compute contrastivity of the model for different classes based on the outcome of STG-Grad-CAM. In the cross-person setting, we observe better contrastivity than the cross-view setting.

***Index Terms*—** Grad-CAM, Skeleton, Graph based methods, Spatio-temporal graph neural network, NTU-RGB-D.

## 1. INTRODUCTION

Convolutional neural networks (CNNs) [1] have led to significant advances in a variety of applications, including computer vision, image and video understanding [2] [3], where input data have a regular structure (e.g., regular grids for image data). More recently, graph convolutional neural networks (GCNNs) [4], [5] have been proposed to deal with data having an irregular structure, such as social networks, skeleton based motion capture data (Mocap), chemistry molecular data [6], etc. GCNNs have also provided excellent results, following the same end-to-end strategy as CNNs, i.e., learning many filters from data.

In applications such as human activity understanding from skeleton data [7], or road network data for traffic forecasting [8], temporal information also needs to be incorporated into the learning process. This has led to the development of spatio-temporal graph convolutional networks (STGCN) [9], [10], which can automatically capture the patterns embedded in the spatial configuration of the nodes as well as in their temporal dynamics. STGCNs have shown superior performance for human activity understanding using skeleton data [9], [11] [12], [7]. Authors have attempted to explain STGCNs based on their feature maps [12]. However, there is only very limited interpretability in these results, which makes it challenging to adapt them for use in a similar but different application and to make informed decisions to improve the performance of these systems.

Researchers have been exploring techniques for improving interpretability of CNNs for a number of years [13] [14], [15] [16], [17]. Recently, these ideas have been extended to GCNNs. For example, Pope et al. [18] proposed some methods to understand GNNs in molecular structure understanding and for scene graphs. Authors in[19] designed an interpretable brain network based on the GNN framework for disease prediction. However, none of these methods deals with the temporal aspects of the problem. In order to achieve better interpretation of STGCNs, it is necessary to capture the importance of both the spatial (vertex domain) and temporal features.

Inspired by the explainability work on CNNs [16] and GCNNs [20], we introduce methods to interpret the output of STGCNs. We extend Gradient-weighted Class Activation Mapping (Grad-CAM) [16] to spatio-temporal graph convolution data and use it to interpret the predictions made by STGCNs. Our proposed method, spatio temporal graph Grad-CAM (STG-Grad-CAM), is generic and it can be used for any spatio-temporal graph topology. Grad-CAM techniques have the potential to be more helpful for spatio-temporal graphs than for images, because it may be hard for humans to intuitively determine the relative importance of different graph nodes for specific classification tasks.

The spatio temporal graph convolution we consider here is a separable transform, so that both the spatial and temporal convolution retains separate localized information and STGCN retains localized information at any point in space and time. STG-Grad-CAM uses the gradient of any query (action) class flowing to the final convolutional layer which is a function of the graph adjacency matrix and the previous layer's output. A spatio-temporal pooling is applied to the gradients to generate a weight vector for the features. The weighted average of these ST features produces a heatmap highlighting the most important graph nodes (e.g., activated body joints in an action classification task) along with their most influential dynamic pattern. The proposed method can perform spatio-temporal graph localization in graphical data with a dynamic temporal variation. Further, we use class-wise temporal pooling to generate node importance maps and find which nodes (body joints for a skeleton graph) are contributing to the model's decisions for a particular task. We extend the concept of faithfulness[16] to evaluate explainability of STG-Grad-CAM. Additionally, we compute contrastivity [18] among classes based on the outcome of STG-Grad-CAM to assess the model's performance.

To evaluate this proposed approach, we consider the problem of human activity recognition using a spatio-temporal skeleton graph. We first create a joint temporal importance map for this specific task that explains which body joints and their temporal dynamics are responsible for a specific action. We further find the dominant structure of the spatial graph for a particular action. We present an interpretability study of this skeleton-based STGCN model in different

experimental settings: 1) changing the train-test protocols to incorporate cross-person (persons in the test set are not the same as in the training set) and cross camera view (similarly camera angles are different in training and testing sets) [21], 2) varying the model depth. Our specific contributions in this work are:

- Adapting Grad-CAM to spatio-temporal graph convolution.
- Presenting interpretability of STGCN in a skeleton-based activity recognition task.
- Presenting a quantitative metric characterizing 'faithfulness' of STG-Grad-CAM and explaining model's 'contrastivity' among classes based on STG-Grad-CAM.

While our results are for a skeleton-based activity recognition task, our proposed method is general and can be used to understand any spatio-temporal graph convolution based task.

## 2. PROPOSED METHOD

### 2.1. Spatio-temporal graph convolutional network

STGCN is a general framework to process structured time-series where general spatio-temporal sequence learning tasks are involved. A spatio-temporal convolutional block in the network combines graph convolutions and temporal convolutions, which can capture the most useful spatial and temporal features. In the case of activity recognition based on skeleton data, each body joint of the skeleton represents a node of the skeleton graph, where the graph signal is the 3-D position of the node. As an example, we work with a spatio-temporal skeleton graph [9], [22] with i) spatial edges which connect the physical joints of a skeleton within a frame, and ii) temporal edges, which connect the same skeleton graph nodes in two consecutive frames.

Let an attributed spatial graph with $N$ nodes be defined by its node $\mathbf{X} \in \mathbb{R}^{d_{in} \times N}$ attributes and its adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, here, $d_{in}$ is the length of input feature. The normalized adjacency matrix is denoted as $\boldsymbol{\mathcal{A}} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is the degree matrix and $\mathbf{I}$ is added to incorporate a self-loop to each node. The spatio-temporal input signal is $\mathbf{X} \in \mathbb{R}^{d_{in} \times N \times T}$ where $d_{in}$, $N$ and $T$ represent the number of channels, number of joints and the temporal length of the data, respectively.Spatio-temporal graph convolutional is computed using:

$$\mathbf{F}^l = \sigma((\mathbf{D}^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{D}^{-\frac{1}{2}}) \odot \mathbf{Q}\mathbf{F}^{l-1}\mathbf{W}^l), \tag{1}$$

where $\mathbf{F}^l$ is the convolutional feature at the $l^{th}$ layer and $\mathbf{F}^0 = \mathbf{X}$. $\mathbf{W}^l$ represents the stacked weight vectors of multiple output channels, the temporal filters are shared by all nodes, $\mathbf{Q}$ learns the spatial edge weights of the graph, $\odot$ denotes the element-wise dot product and $\sigma$ represents the element-wise nonlinear activation function.

Multiple layers of spatio-temporal graph convolution units are used on the input data. We use a configuration of skeleton-based activity recognition using STGCN similar to that in [9]. The STGCN model comprises 10 layers of spatial temporal graph convolution operators (STGCN units) followed by global average pooling, fully connected layer, and softmax.

### 2.2. STG-Grad-CAM

Similar to regular convolutional layers, spatio-temporal graph convolution layers also retain spatially and temporally localized information, which is lost in fully-connected layers (FCL). Thus, the last convolutional layer before FCL contains optimal high-level spatio-temporal information. The neurons in these layers are responsible for the class-specific spatio-temporal importance of the skeleton joints in the activity sequence data. The gradient information flowing into the last spatio-temporal graph convolutional layer of the STGCN is used in Grad-CAM to compute the importance of each neuron for a particular decision of interest. Let the $k^{th}$ graph convolutional feature map at layer $l$ be defined as:

$$\mathbf{F}^l_k(\mathbf{X}, \mathbf{A}) = \sigma(\tilde{\mathbf{A}}\mathbf{F}^{l-1}(\mathbf{X}, \mathbf{A})\mathbf{W}^l_k). \tag{2}$$

Here, the $k^{th}$ feature at the $l^{th}$ layer is denoted by $\mathbf{F}^l_{k,n,t}$ for node $n$ and time $t$ and $\tilde{\mathbf{A}} = (\mathbf{D}^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{D}^{-\frac{1}{2}}) \odot \mathbf{Q}$. Then, the global average pooling (GAP) feature after the last STG-convolutional layer ($L$) is computed as $\boldsymbol{e}_k = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbf{F}^L_{k,n,t}$. The class score is calculated as $y_c = \sum_k \boldsymbol{w}^c_k \boldsymbol{e}_k$. The class activation maps (CAM) can be calculated as

$$H^c_{STG-CAM} = \|\mathrm{ReLU}\left(\sum_k \boldsymbol{w}^c_k \mathbf{F}^L_k\right)\|.$$

Using this notation, we extend the Grad-CAM method for STGCNs and name it as Spatio-temporal graph grad cam (STG-Grad-CAM). **Gradient-based** heatmaps over nodes $n$ and time frames $t$ are:

$$H^c_{Gradient} = \|\mathrm{ReLU}\left(\frac{\delta y^c}{\delta \mathbf{X}_{n,t}}\right)\|. \tag{3}$$

STG-Grad-CAM's class specific weights for class $c$ at layer $l$ and for feature $k$ are calculated by:

$$\boldsymbol{\alpha}^{c,l}_k = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{\delta y^c}{\delta \mathbf{F}^l_{k,n,t}}. \tag{4}$$

The final heatmap calculated from layer $l$ is

$$H^{c,l}_{ST} = \mathrm{ReLU}\left(\sum_k \boldsymbol{\alpha}^{c,l}_k \mathbf{F}^l_k\right). \tag{5}$$

STG-Grad-CAM enables us to generate class-specific heatmaps for different layers of the network. Using these heatmaps, we can understand the importance of the different body joints over time for a specific task. In other words, which joint is responsible for selecting a specific label for a specific action. To create the body-joint importance map ($\psi$) for each action class, we compute the STG-Grad-CAM for each class and take an average over all the correctly classified data-points. Let $M^C$ be the number of correctly classified data-points for an action class $C$. Then, the joint importance map is then defined by

$$\psi^C_n = \frac{1}{M^C}\sum_{m=1}^{M^C}\frac{1}{T^C_m}\sum_{t=1}^{T^C_m}H^{C,L,n,t}_{ST} \tag{6}$$

Here, $T^C_m$ denotes the number of frames in the $m^{th}$ correctly classified data-point of class $C$. In the end, we normalize the joint importance map $\psi^C$ for all joints within an action between (0,1).

## 3. EXPERIMENTAL RESULTS

### 3.1. Dataset

In this experiment, we use the NTU-RGB+D [21] datasetThis has 56,000 data points, including 60 actions. 40 volunteers performed

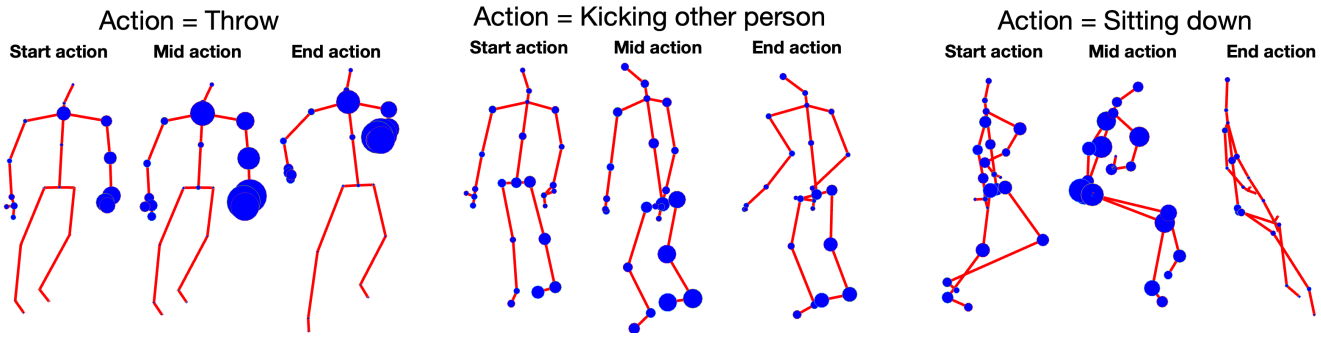


**Fig. 1**: Actions at different time points with their spatio temporal joint importance: a bigger node size represents higher importance.

these actions while this was captured in three-camera placed at different angles. Each data point contains the temporal sequence of 3-D locations of 25 joints of the participants. Following the recommended train-test protocol, we also evaluate our system in 1) cross-subject (*xsub*) and 2) cross-view (*xview*).

### 3.2. Results

To evaluate the performance of STG-Grad-CAM, we consider a human action classification task using skeleton data. The 10-layer STGCN model [9] is used for classification, which can achieve $82.1\%$ accuracy in *xsub* setting and $88.5\%$ accuracy in *xview* setting. Now, to understand the STG-convolutional model, we apply STG-Grad-CAM and generate a class-specific heatmap using (5) for all the data points. Fig. 1 shows the joint importance for three actions at different points of time. The bigger size of the node of the skeleton graph denotes the higher importance of the body joint at that point in time. For the first action *throw* in which mostly the hands are involved, we see more activation in the hand joints, specifically in the right hand. Action 'Kicking other person' mostly involves lower body parts, and in the figure also, we see higher importance in the leg and the back region. Therefore, we can see the dominant substructure of the graph for this action. Note that the important changes over time, and it is highest in the middle of the action. Action *sitting down* usually requires all the joints with more involvement of the center of a human body which is clearly seen in fig. 1. Interestingly, at the end of the task, all the joints are not highly activated.

Now, to understand the importance of human body joints for a particular task, we compute joint importance value using (6) and take an average overall data points. Fig. 2 shows the joint importance map for all the actions. Interestingly, actions starting from #49 involve lower body than upper body, which is clearly seen in the figure. This indicates the influential joints behind STGCN's prediction for each class. This result is generated in the *xsub* setting.

In *xview* setting, a given action for one participant is present in both the training and test set, which are captured at a different camera angle. Naturally, STGCN achieves better accuracy in *xview* setting. Fig. 3 shows the joint importance map for all actions in *xview* setting. We can see that there is significant importance in the lower body region when the action predominantly involves upper body or hands, for example, nausea, wear jacket, writing.

**Faithfulness** : To characterize the faithfulness of the proposed STG-Grad-CAM, we mask the data from nodes that are highlighted by STG-Grad-CAM and then measure the change in the accuracy, where the mask takes a value in the interval $(0, 1)$. Table 1 shows the accuracy achieved by the model for a varying amount of masking in the highlighted nodes indicated using STG-Grad-CAM. It is expected that with a higher amount of masking, we get lower accuracy and vice versa, which is evident in the table. Table 2 shows the accuracy achieved by the model for a varying amount of masking in the nodes identified as less important STG-Grad-CAM. Comparing Table 1 and Table 2 we can see that masking less -important nodes does not affect the performance significantly as compared to masking the important nodes. This provides further justification of the correctness of the node importance map.

Let the masking percentage be $m \in [0, 1]$ and the corresponding accuracy be $\alpha_m$, where $\alpha_0$ denotes the accuracy without masking. Then, faithfulness is computed using:

$$\beta_{faithful} = \sum_i m_i \left( \frac{|\alpha_0 - \alpha_{m_i}|}{\alpha_0} \right). \tag{7}$$

Table 3 shows $\beta_{faithful}$ for *xsub* and *xview* setting. The values of $\beta_{faithful}$ in *xsub* and *xview* settings are consistent with their accuracy value.

**Contrastivity**: Inspired by the concept of Contrastivity mentioned in [18] in a binary classification problem, we define the contrastivity ($\beta_{contrast}$) for STG-Grad-CAM in this case for a multi-class classification problem. Contrastivity is defined to characterize the intuition that class-specific features highlighted by an explanation method should differ between classes. Once we have the node importance map ($\psi$), we find the correlation [23] between the node importance vector $\psi^C$ for a class $C$ with all other classes. Then contrastivity ($\beta_{contrast}$) is computed by taking an average of them and then subtracting it from 1. Higher value of $\beta_{contrast}$ represents better contrastivity .

Table 3 shows $\beta_{contrast}$ for *xsub* and *xview* setting. Though we achieve better accuracy in *xview* setting, *xsub* achieves better contrastivity showing better separation in feature space among classes. It is possible that *xview* model is performing better in terms of accuracy because of the subjective bias present in the data and it is not solely concentrating on the action-specific joint dynamics.

**Table 1**: Accuracy achieved by STGCN for NTU-RGB-D for different amount of masking of the data of the important nodes

| Masking (in %) | No masking | 10% | 50% | 90% |
|---|---|---|---|---|
| *xsub* - Accuracy(%) | 81.5 | 80.3 | 64.52 | 20.59 |
| *xview* - Accuracy(%) | 88.3 | 88.1 | 64.31 | 14.82 |

We also studied the same STGCN network with fewer STGCN layers. We removed the $3^{rd}$, $6^{th}$ and $9^{th}$ layer (chosen randomly)

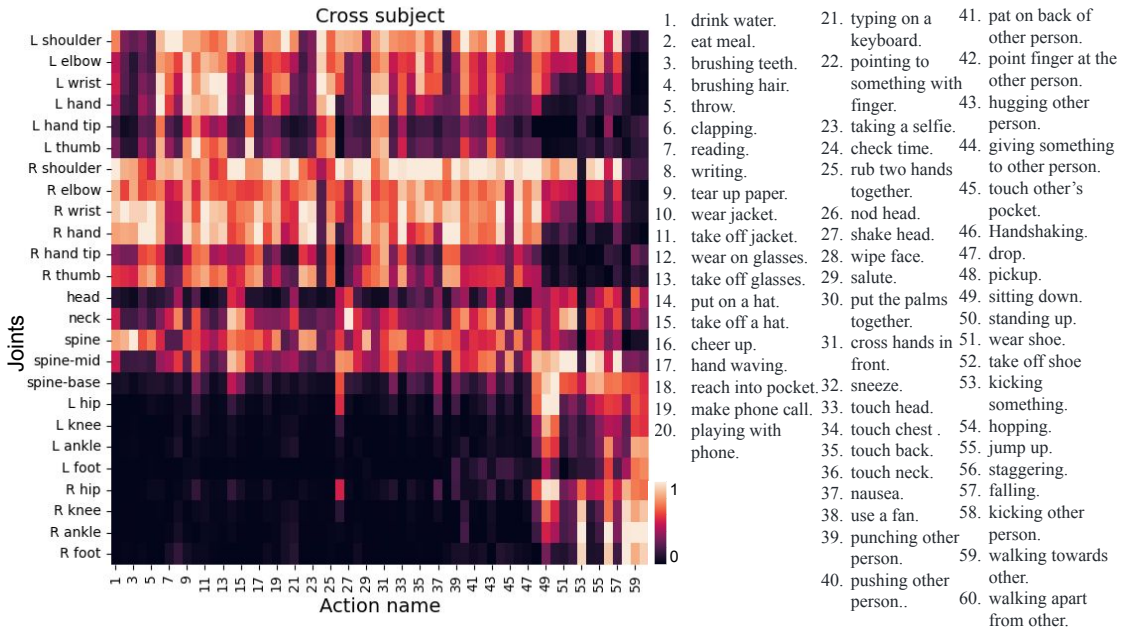


**Fig. 2**: Joint importance map for all actions, In the Y tick labels 'L' : 'Left', 'R' : 'Right'. Action names in the X-axis is given in the right of the figure. This figure is generated in x-sub setting.

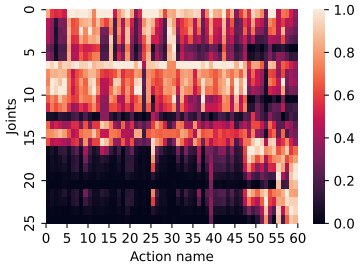


**Fig. 3**: Joint importance map in *xview* setting.The action and joint order in X and Y axis respectively follow the order shown in fig 2.

**Table 2**: Accuracy achieved by STGCN for NTU-RGB-D for different amount of masking of the data of the non-important nodes

| Masking (in %) | No masking | 10% | 50% | 90% |
|---|---|---|---|---|
| Xsub - Accuracy(%) | 81.5 | 80.57 | 72.5 | 66.5 |
| Xview - Accuracy(%) | 88.3 | 88.1 | 81.09 | 66.81 |

**Table 3**: Fathfullness and contrastivity in *xview* and *xsub* settings

| Experimental protocol | *xview* | *xsub* |
|---|---|---|
| $\beta_{contrast}$ | 0.51 | 0.54 |
| $\beta_{faithful}$ | 0.89 | 0.74 |

with corresponding number of output filters 64, 128 and 256. This STGCN model with 7 layers (STGCN 7) achieved 70.2% accuracy for NTU-RGB-D dataset in an *xsub* setting. Fig. 4 shows the joint importance map for this network. Surprisingly, for all the actions, only the hand region is mostly influencing the decision. Naturally, the misclassification is higher.

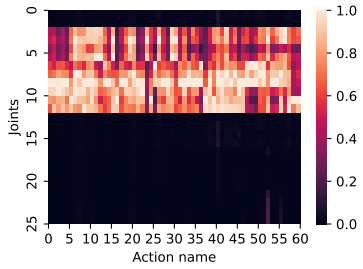


**Fig. 4**: Joint importance map for STGCN-7. The action and joint order in X and Y axis respectively follow the order shown in fig 2.

## 4. CONCLUSION

In this paper, we present an interpretability method STG-Grad-CAM to understand STGCN. STGCN can handle any spatially structured data varying over time, and STG-Grad-CAM can identify the influential nodes behind the model's decision. We generate a joint-time activation heatmap to understand how these joints and their temporal dynamics contribute to the model's decision. Further, we create a joint activation map for each action. We analyze interpretability using STG-Grad-CAM in different experimental settings. We evaluate the explanation of STG-Grad-CAM using quantitative metrics, faithfulness. We observe higher faithfulness in the *xview* setting. Additionally, we show the model's contrastivity among classes based on the outcome of STG-Grad-CAM and find higher contrastivity in *xsub* setting. We study how the change in the model depth affects the performance.

## 5. REFERENCES


[1] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, 2015.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams, "Convolutional networks on graphs for learning molecular fingerprints," *arXiv preprint arXiv:1509.09292*, 2015.

[5] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[6] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande, "Moleculenet: a benchmark for molecular machine learning," *Chemical science*, vol. 9, no. 2, pp. 513–530, 2018.

[7] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 143–152.

[8] Bing Yu, Haoteng Yin, and Zhanxing Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," *arXiv preprint arXiv:1709.04875*, 2017.

[9] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[10] Chao Pan, Siheng Chen, and Antonio Ortega, "Spatio-temporal graph scattering transform," *arXiv preprint arXiv:2012.03363*, 2020.

[11] Pratyusha Das and Antonio Ortega, "Symmetric sub-graph spatio-temporal graph convolution and its application in complex activity recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3215–3219.

[12] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3595–3603.

[13] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[14] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, John E Hopcroft, et al., "Convergent learning: Do different neural networks learn the same representations?," in *FE@ NIPS*, 2015, pp. 196–212.

[15] Aravindh Mahendran and Andrea Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.

[16] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[17] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu, "Interpretable convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8827–8836.

[18] Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann, "Explainability methods for graph convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10772–10781.

[19] Hejie Cui, Wei Dai, Yanqiao Zhu, Xiaoxiao Li, Lifang He, and Carl Yang, "Brainnnexplainer: An interpretable graph neural network framework for brain network based disease analysis," *arXiv preprint arXiv:2107.05097*, 2021.

[20] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[21] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[22] Jiun-Yu Kao, Antonio Ortega, Dong Tian, Hassan Mansour, and Anthony Vetro, "Graph based skeleton modeling for human activity analysis," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 2025–2029.

[23] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*, pp. 1–4. Springer, 2009.