USC Viterbi School of Engineering

Ming Hsieh Department of Electrical Engineering

Gradient-weighted Class Activation Mapping for spatio temporal graph convolutional network

Pratyusha Das, Antonio Ortega

University of Southern California (USA)

Introduction

- **Motivation:**
  - Spatio-temporal Graph Convolutional Neural Network.
  - Captures simultaneously the spatial correlation and the temporal pattern in the data
  - Lacks interpretability
  - Unknown reason behind their prediction
- **Contribution:**
  - Spatio-temporal Graph Grad CAM
  - Gradient based class activation maps for STGCN

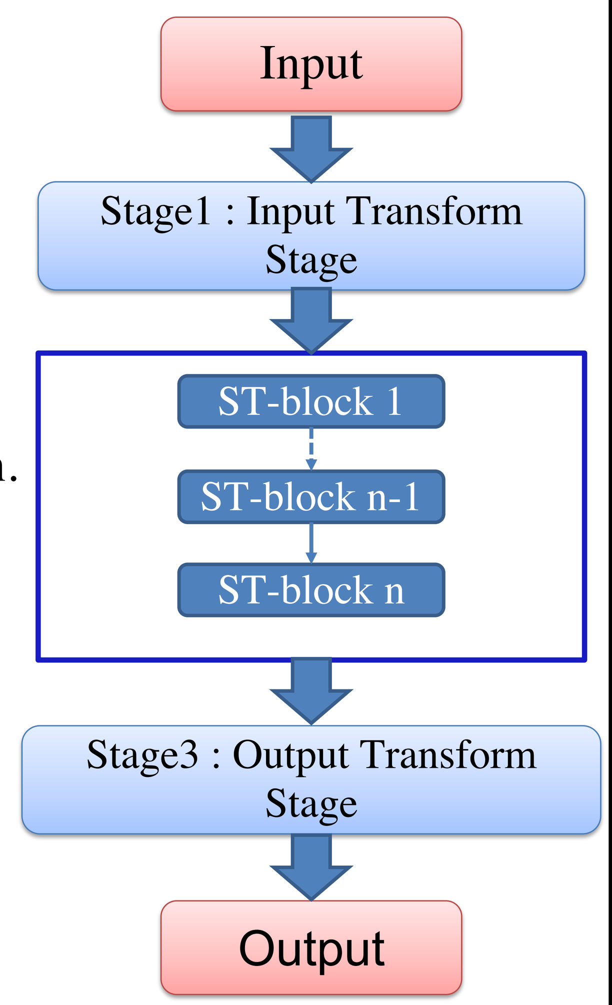
Preliminaries

STGCN [1]

In each layer, ST-GCN is computed as

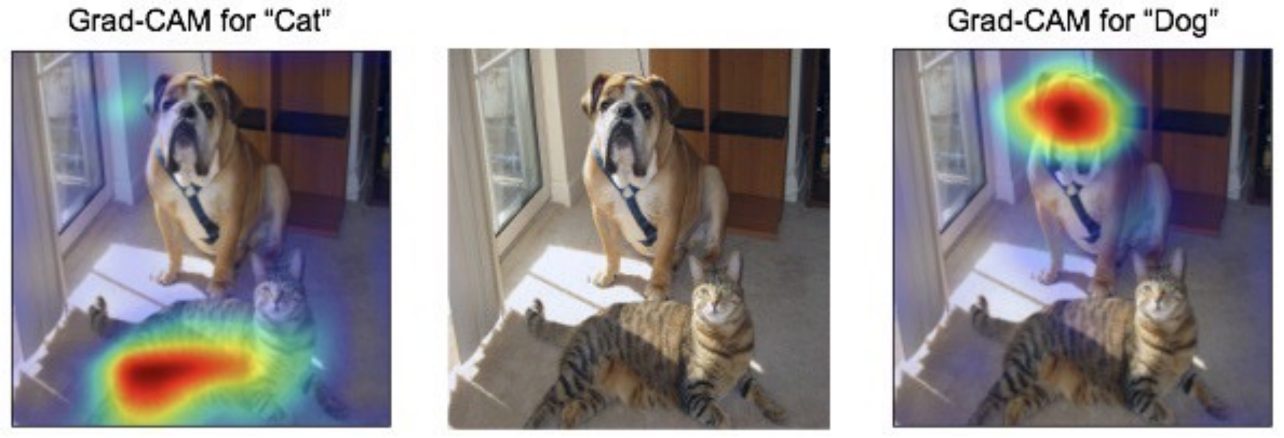
$$x_{out} = (\tilde{A} \odot Q) x_{in} W$$

- x_{in} : input feature map of size (C, V, T)
- C : Channel, V : Vertices, T : Temporal length.
- x_{out} : output feature map.
- $\tilde{A}$: adjacency matrix of spatial graph.
- Q : edge weight matrix of the spatial graph.
- W : stacked weight vector of the multiple output channel.



Grad-CAM [2]

- CNN uses 2-D spatial convolutional filter.
- Gradient based class activation maps (Grad-CAM) detects the spatial image important region for the network



Spatio-emporal graph Grad CAM

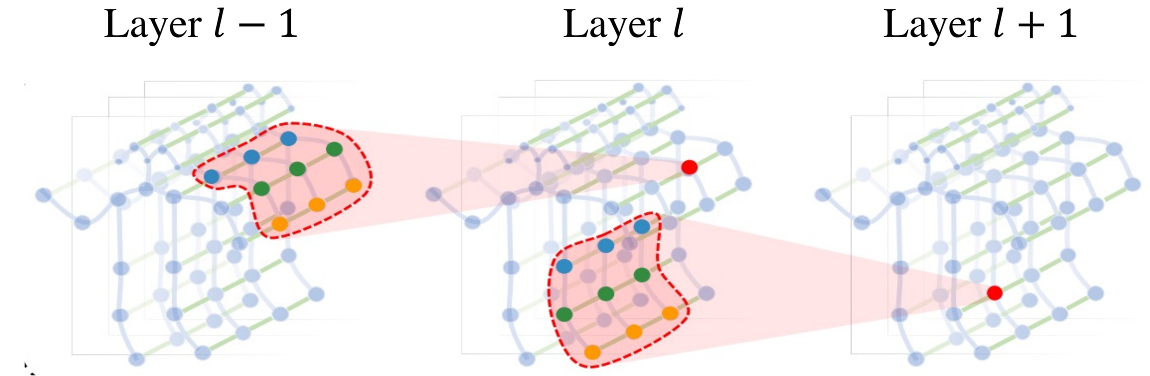
- Adjacency matrix as filter

$$((A+I)x)_i = \sum_{j \in \mathcal{N}(i)} w_{i,j} x_j + w_{i,i} x_i$$

- The spatio–temporal GCNN is computed as

$$F_k^l(X, A) = \sigma\left(\tilde{A} F^{l-1}(X, A) W_k^l\right)$$

- Here, normalized adjacency matrix $\tilde{A}$, F_k^l represents the k-th feature at the l-th layer.
- Features are localized spatially and temporally.



- The heatmap $H_{ST}^{c,l}$ for joint-time importance for c^{th} class and l^{th} layer is computed using

$$H_{ST}^{c,l} = \mathrm{ReLU}\left(\sum_k \alpha_k^{c,l} F_k^l\right)$$

- $\alpha_k^{c,l}$: Weights for k-th feature, F_k^l : k-th feature map.
- ReLU considers the positive value which contributed to the final decision
- Weights are calculated based on the gradients

$$\alpha_k^{c,l} = \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \frac{\partial y^c}{\partial F_{k,n,t}^l}$$

- Here, y^c represents c-th class score.

Evaluation

Faithfulness : Measures the impact of occlusions to the graph nodes

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

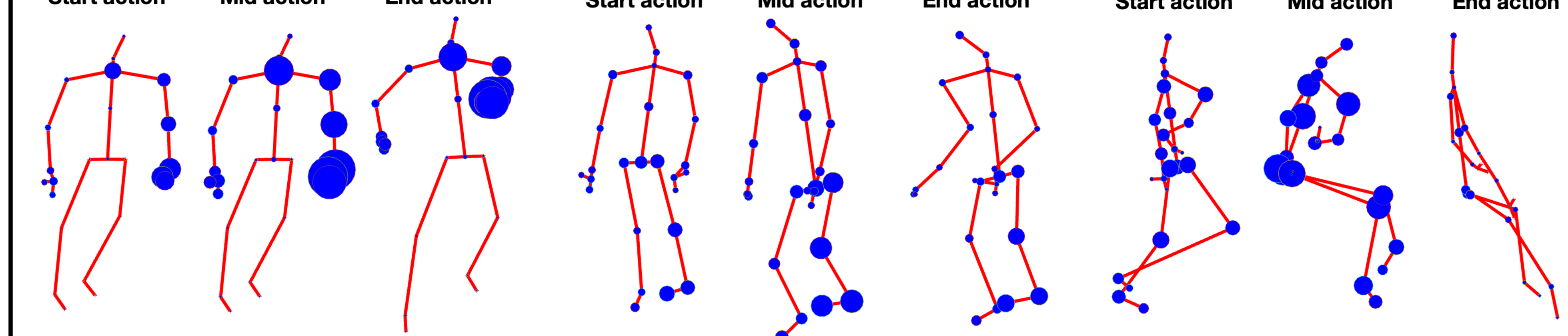
masking percentage m, corresponding accuracy α_{m_i} , α_0 is accuracy without masking

- Achieved $\beta_{faithful} = 89\%$

Experiments and Results

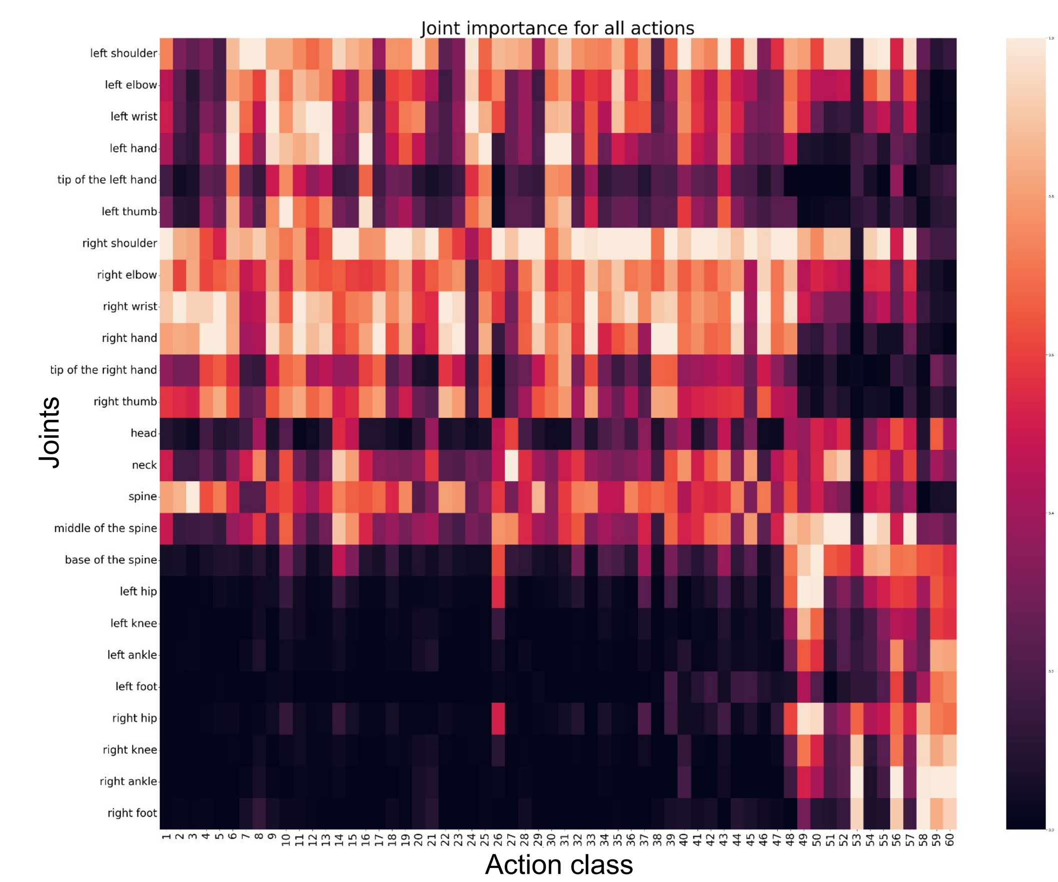
- **Task: Skeleton based activity recognition**
- **Model: STGCN** 10 layers of STGCN, Trained on NTU-RGB dataset, No of classes 60, Accuracy 81.5%

Fig. Actions at different time points with their spatio temporal joint importance



*Bigger size of the node represents higher importance of that joint.

Joint importance map → $$\psi_n^C = \frac{1}{M^C} \sum_{m=1}^{M^C} \frac{1}{T_m^C} \sum_{t=1}^{T_m^C} H_{ST}^{C,L,n,t}$$



1. drink water.
2. eat meal/snack.
3. brushing teeth.
4. brushing hair.
5. throw.
6. clapping.
7. reading.
8. writing.
9. tear up paper.
10. wear jacket.
11. take off jacket.
12. wear on glasses.
13. take off glasses.
14. put on a hat/cap.
15. take off a hat/cap.
16. cheer up.
17. hand waving.
18. reach into pocket.
19. make a phone call/answer phone.
20. playing with phone/tablet.
21. typing on a keyboard.
22. pointing to something with finger.
23. taking a selfie.
24. check time (from watch).
25. rub two hands together.
26. nod head/bow.
27. shake head.
28. wipe face.
29. salute.
30. put the palms together.
31. cross hands in front (say stop).
32. sneeze/cough.
33. touch head (headache).
34. touch chest (stomachache/heart pain).
35. touch back (backache).
36. touch neck (neckache).
37. nausea or vomiting condition.
38. use a fan (with hand or paper)/feeling warm.
39. punching/slapping other person.
40. pushing other person.
41. pat on back of other person.
42. point finger at the other person.
43. hugging other person.
44. giving something to other person.
45. touch other person's pocket.
46. Handshaking.
47. drop.
48. pickup.
49. sitting down.
50. standing up (from sitting position).
51. wear a shoe.
52. take off a shoe.
53. kicking something.
54. hopping (one foot jumping).
55. jump up.
56. staggering.
57. falling.
58. kicking other person.
59. walking towards each other.
60. walking apart from each other.

Accuracy achieved by STGCN for different amount of masking of the data of the joints based on their importance

Node type		Important nodes			Non-important nodes		
Masking (in %)	No masking	10%	50%	90%	10%	50%	90%
Cross subject -Accuracy (%)	81.5	80.3	64.52	20.59	80.3	72.5	66.5
Cross Camera view -Accuracy (%)	88.3	88.1	64.31	14.82	88.1	81.09	66.81

Future work

- Interpretability of STGCN
- Generic method – change the graph to implement if for other applications.

[1] Yan, S, et al.. "Spatial temporal graph convolutional networks for skeleton-based action recognition." *AAAI* 2018.
[2] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *ICCV* 2017..

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA