Actional-Structual Graph Convolutional Networks for Skeleton-based Action Recognition

Miaoran Chen

Part Ⅰ Basic information

1. **Title:** Actional-Structual Graph Convolutional Networks for Skeleton-based Action Recognition
2. **Authors:** Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, Qi Tian.

Part Ⅱ Paper structure

1. **Introduction**

Skeleton data, representing dynamic 3D joint positions, have been shown to be effective in action representation. While previous studies are mostly based on fixed skeleton graphs, only capturing local physical dependencies, not accurate enough for action recognition. To solve the problem of encoding and processing topological data, GCN has been proposed. Since skeleton data is topological, recent methods such as ST-GCN (extracts the features of joints directly connected via bones) has been proved effective. However, structurally distant bones which may cover key patterns of actions are ignored in ST-GCN. This paper focus on how to improve ST-GCN.

1. **Problem Description**

Recognizing human action based on skeleton data.

1. **The Proposed Algorithm**
2. A-links

A-links (Actional links) are activated by actions and might exist between arbitrary pair of joint. It represents corresponding dependencies for various actions. AIM (A-link inference module) is a trainable module and is used to infer A-links from actions automatically. The encoder of AIM produces A-links by propagating information between joints and links iteratively to learn link features. The decoder predicts future joint positions based on the inferred A-links.

1. S-links

S-links (Structural links) are used to represent long-range links in skeleton data. Like ST-GCN, this paper uses the high-order polynomial to indicate S-links. And fusing S-links with traditional convolutional kernel to increase receptive field.

1. Connection

Convex combination is used to integrate actional structural convolution and structural convolution. The proportion of the two parts could be changed.

1. **Experiments and Discussions**
2. Dataset

NTU-RGB+D: 568000 skeleton action sequences, 60 classes.

Kinetics: Kinetics is a representative dataset for human action analysis. Skeleton data is obtained by estimating joint locations on certain pixels.

1. Result

When using both S-links and A-links, the algorithm achieves best performance. An additional future pose prediction head captures more detailed patterns through self-supervision.

1. **Conclusion**

The AS-GCN proposed a successful solution to improve the performance of ST-GCN. Also, it shows

Promising results for future pose prediction.

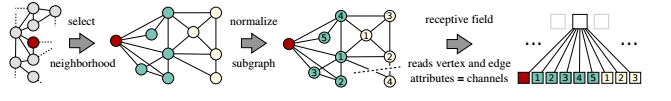
Part Ⅳ Perspective

I regard this paper as a promotion of [Spatial Temporal Graph Convolutional Networks for Skeleton Based Action Recognition](https://arxiv.org/abs/1801.07455). On the basis of ST-GCN, this paper add a parameter: A-link to optimize the performance.

While I was reading, 2 key points attracted me: the input of network, GCN.

While I was doing investigation of human action recognition, I noticed that optic flow is mainly an indispensable input(Action recogniton.docx). There are 2 main research directions in human action recognition: network structure and network connection. Within the problem connection, some attempts to find out an input to replace optic flow graph had been carried out. So I was surprised to acknowledge that researchers used skeleton data as input. In my word, skeleton data in human action recognition is similar to the landmark in facial recognition. Both of them apply the philosophy of simplification and modelling.

While CNN is indeed effective in many computer vision problems, it could not process data with Non Euclidean Structure. However, data generated from reality always has such structure. GCN is used to solve this problem. It advanced the encoding method and was more adaptive to real data. In my opinion, the necessity of GCN is correspond to our requirement or demand to extract spatial features from topological graph. CNN is relatively restrained because the size of convolutional kernel is constant. Broadly speaking, topology relation could be established to any kind of data in normed space(spectral clustering is an example for employing this thought). Thus I think GCN has enormous development space in the future. I learned the basic theory of GCN. There are two methods to extract spatial features from a topological graph: vertex domain and spectral domain. The main idea of vertex domain is to extract adjacency characteristic of each vertex. However, it is hard to determine the receptive field and process neighbors’ features. The picture below shows the main step of vertex domain([Learning Convolutional Neural Networks for Graphs](http://proceedings.mlr.press/v48/niepert16.pdf)).



Spectral domain has more widespread use. It combines 3 main parts: Fourier Transformation on graph, graph convolution, deep learning. The major elements of spectral domain is Laplacian Matrix(Combinatorial Laplacian, Symmetric normalized Laplacian, Random walk normalized Laplacian). Some points I did not figure out on this problem. I would look into it in the next week. Many attempts to use GCN to solve problems have proceeded. [Multi-Label Image Recognition with Graph Convolutional Networks](https://arxiv.org/abs/1904.03582)(2019), [MotifNet: a motif-based Graph Convolutional Network for directed graphs](https://arxiv.org/abs/1802.01572)(2018) are relatively well known examples.

For I am new to human action recognition, give my opinions prematurely seems not wise. I would read more papers and do more research of this field before give my comments.