Multi-scale Convolution for Spatio-Temporal Modeling 4D Point Cloud

Jinglue Hang[a], Qiang Liu[a]

[a]*Dalian University of Technology,*

Abstract

The work of the 2nd place in the HOI4D Challenge 2023 (Action Segmentation Track).

1. Introduction

The HOI4D[Liu et al. (2022)] action segmentation task gives each frame of the point cloud in the point cloud video an action category label and the output is the action category of each frame in the video. Our work builds upon the modification of the PPTR [Wen et al. (2022)] model provided by the sponsor. Our network achieved an accuracy of 0.843 on the test set. Note that we do not employ external training data, only modify the network structure. The main contributions of our work are as follows:

(1) Drawing inspiration from the Multi-scale grouping technique in Pointnet++ [Qi et al. (2017)], we propose a multi-scale convolution block upon the 4D backbone of PPTR. This module utilizes 4 different spatial radii for 4D convolution to extract features, these features are added resulting in improvement.

(2) Inspired by the STSA (Spatial-Temporal Self-Attention) module in PSTT [Wei et al. (2022)], we enhance the transformer module by incorporating residual connections between the self-attention calculation results and the input features. This helps form the final input features and further improves the network's effectiveness.

(3) During the training process, we employ several tricks specifically for the transformer module, we introduce a dropout layer (0.1 ratio) and increase transformer block depth (from 5 to 6) to enhance the network's performance.

2. Method

2.1. multi-scale convolution

PointNet++ is a remarkable work that focuses on feature extraction of point cloud in classification and segmentation tasks. It addresses the challenge of handling irregularly distributed point clouds by introducing the Multi-scale Grouping module, which enhances the effectiveness of classifying and segmenting static point clouds. In the context of 4D action segmentation, extracting point cloud features is also a crucial step. Therefore, our first improvement involves enhancing the network's backbone of PPTR.

The 4D backbone of PPTR defines spatio-temporal convolution, which effectively captures features from 4D point cloud sequences. Spatial convolution involves using the ball query method in the temporal domain to extract point clouds from different frames for feature extraction. Building upon the Multi-scale Grouping method of PointNet++, we perform feature extraction using the same ball query with various spherical radii, denoted as r. Notably, unlike PointNet++, we use addition instead of concatenation for feature fusion. Through extensive experimentation, we have found that addition achieves superior feature extraction compared to concatenation, resulting in higher accuracy on the test set.

Specifically, we conduct four rounds of feature convolution using the original radius (r=0.9), r multiplied by 2, r multiplied by 4, and r multiplied by 6, respectively. We then add the resulting features. This modified backbone significantly enhances the network's performance by around 2 percentage points. However, it does come with a drawback of approximately a 2GB increase in GPU memory usage.

2.2. transformer block

Since the HOI4D dataset is a sequential sequence, significant improvements have been made to the transformer framework in PPTR, which motivates us to enhance the transformer block. PSTT [Wei et al. (2022)] suggests changing the transformer module by addressing the issue of initial weight randomness affecting the consistency between output and input features. To mitigate this, a residual connection is established between the computed self-attention output and the input feature, resulting in the final output feature. Moreover, Layer Normalization is a commonly used technique for expediting the convergence of attention models. It is applied after the residual connection to enhance the network's effectiveness. Therefore, in our network, we enhance the transformer block by incorporating residual connections using the self-attention output and input features, followed by Layer Normalization.

2.3. training tricks

Several common techniques are employed during the training process to address the task. Firstly, we introduced a dropout layer within the transformer block, leading to improved results. Secondly, we increased the depth of the network from the original 5 layers to 6 layers, yielding the best performance. Nevertheless, it is important to note that such approaches utilize limited resources to enhance accuracy and comes with certain limitations.

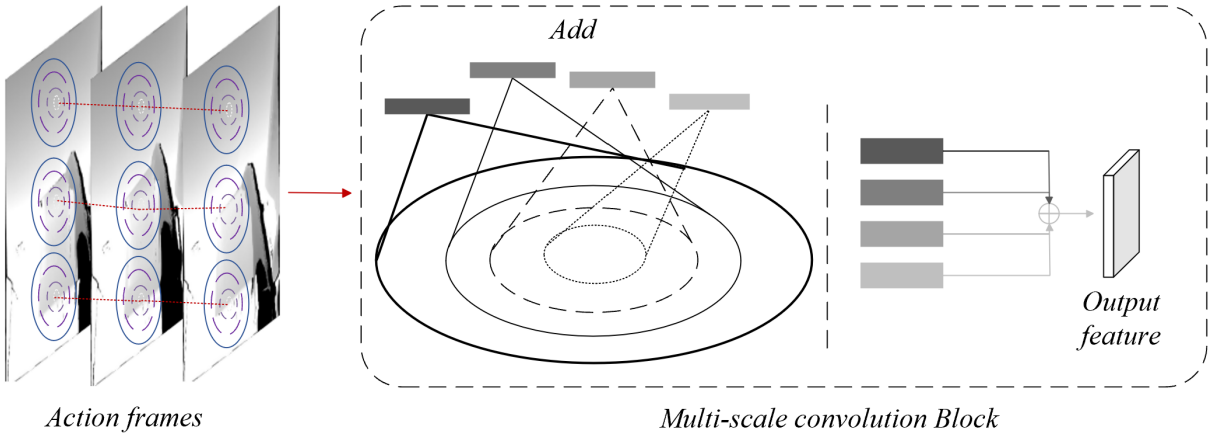

Figure 1: Multi-scale convolution

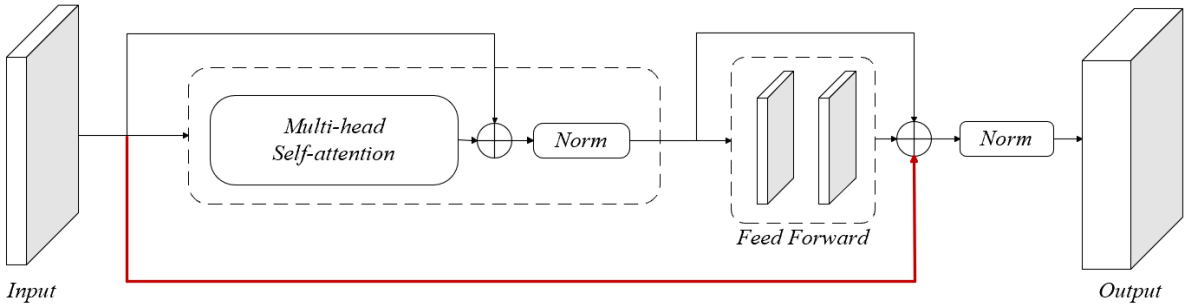

Figure 2: Enhanced transformer block

Method	Accuracy
PPTR (baseline)	0.774
MC+NT	0.829
MC+NT + dropout-layer(0.1)	0.840
MC+NT + dropout-layer(0.1) + depth=6	0.843

Table 1: **MC+NT** means multi-conv + new transformer block mentioned before.

3. Experiments

The experimental results are presented in the table. Based on PPTR, we made network changes. Through several experiments, we employed the original radius (r=0.9), r*2, r*4, and r*6 for four feature convolutions, along with the implementation of a new transformer block. This led to an increased accuracy rate of 0.829, providing strong evidence that combining multi-scale convolution and transformer with residual can significantly enhance accuracy.

Once the network structure is set, we employed various tricks to further improve performance. Initially, we introduced a layer dropout of 0.1. Subsequently, we decided to deepen the depth of the transformer. These adjustments yielded a final network performance of 0.843, representing a huge improvement of baseline.

References

Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L., 2022. Hoi4d: A 4d egocentric dataset for category-level human-object interaction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21013–21022.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems 30.

Wei, Y., Liu, H., Xie, T., Ke, Q., Guo, Y., 2022. Spatial-temporal transformer for 3d point cloud sequences , 1171–1180.

Wen, H., Liu, Y., Huang, J., Duan, B., Yi, L., 2022. Point primitive transformer for long-term 4d point cloud video understanding , 19–35.