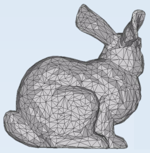


MANet: Multimodal Attention Network based Point- View fusion for 3D Shape Recognition

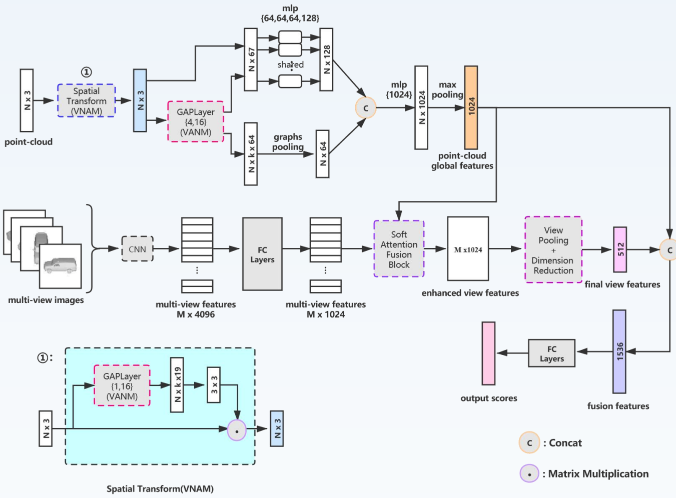
Yaxin Zhao, Jichao Jiao, Ning Li, Zhongliang Deng*

Beijing University of Posts and Telecommunications

Abstract

3D shape recognition has attracted more and more attention as a task of 3D vision research. The proliferation of 3D data encourages various deep learning methods based on 3D data. Now there have been many deep learning models based on point- cloud data or multi-view data alone. However, in the era of big data, integrating data of two different modals to obtain a unified 3D shape descriptor is bound to improve the recognition accuracy. Therefore, this paper proposes a fusion network based on multimodal attention mechanism for 3D shape recognition. Considering the limitations of multi-view data, we introduce a soft attention scheme, which can use the global point-cloud features to filter the multi-view features, and then realize the effective fusion of the two features. More specifically, we obtain the enhanced multi-view features by mining the contribution of each multi-view image to the overall shape recognition, and then fuse the point-cloud features and the enhanced multi-view features to obtain a more discriminative 3D shape descriptor. We have performed relevant experiments on the ModelNet40 dataset, and experimental results verify the effectiveness of our method.

Architecture of the proposed MANet



Basic Concepts

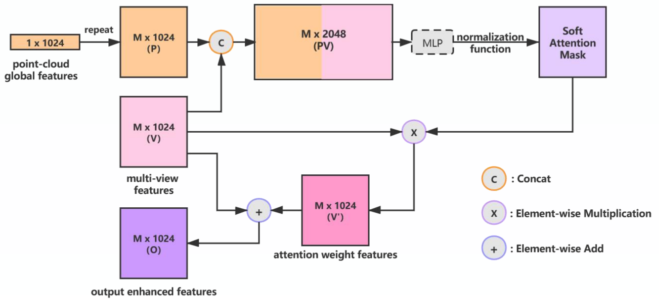
In the attention fusion block, in order to solve the problem that the two types of features are in different feature spaces, we map the global point-cloud features to the subspace of the multi-view features to obtain the features $P = \{P_1, P_2 \cdots, Pm\}$, and then fuse it with the multi-view features $V = \{V_1, V_2 \cdots, Vm\}$ to obtain the fused features $PV = \{I_1, I_2 \cdots, \mathrm{Im}\}$, where m is the number of multi-view images, which we set m=12 in the experiment. Weight coefficients $C(W) = \{W_1, W_2 \cdots, Wm\}$ are generated after normalizing the fused features, i.e.:

$$C(W) = F(MLP(P,V))$$

The normalization function **F(·)** used in this article is the sigmoid function, so the calculation formula of $W_i \in C(W), 1 \le i \le m$ is:

$$W_i = \frac{exp(MLP(P_i,V_i))}{\sum_{k=1}^{m} exp(MLP(P_k,V_k))}, 1 \le i,k \le m$$

Soft Attention Fusion Block



Implementation

Our MANet framework is an end-to-end architecture that introduces the multimodal attention mechanism. Both the point-cloud branch and the multi-view branch have network compatibility. In our experiments, the low-level multi-view features of the 12 images in the multi-view branch are obtained by the method of MVCNN. The overall architecture of the point-cloud branch is based on the classification framework of GAPNet. We have made some improvements, more specifically, we add VNAM module to the part of computing graph features in spatial transformation network and backbone network. In the soft attention fusion block, we use global point-cloud features to perform feature filtering on multi-view features, and use residual connections to obtain the enhanced multi-view features. After that, we fuse the point-cloud features and enhanced multi-view features again to obtain the final overall 3D shape descriptor.

Results

We have compared our MANet with the methods based on volume data, multi-view data, point-cloud data, and fusion data. Our architecture MANet classification accuracy is 93.4%. In the experiments, we find that our model can achieve 100% recognition accuracy on up to 10 categories and also achieve 99% accuracy on objects such as bed and monitor. But it should be pointed out that many models including our MANet exist the problem of having low recognition rate in a certain category. How to make the network distinguish multiple objects accurately will become our future improvement direction.

Conclusion

This paper proposes a fusion network MANet based on the multi-modal attention mechanism for 3D shape recognition. MANet is a new neural network that can use the global point-cloud features to filter multi-view features. In the point-cloud branch, we have improved the GAPNet classification architecture. Here, we propose a VNAM attention module and gain accuracy improvement. In the fusion branch, we design a soft attention fusion block to achieve the processing of multi-view features while fusing the two features. The classification and retrieval experiments on ModelNet40 show the effectiveness of our framework.

References

[1] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 945- 953.

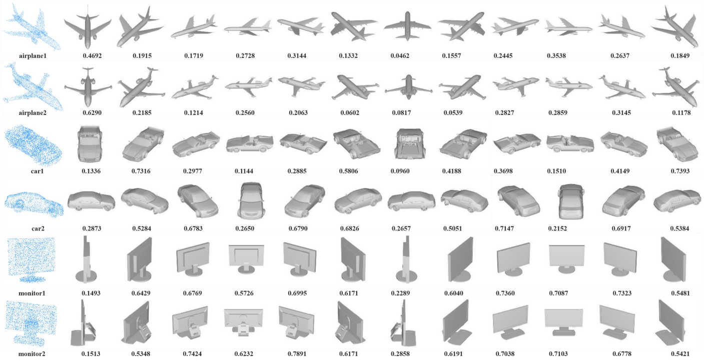
[2] C. Chen, L. Z. Fragonara, and A. Tsourdos, "Gapnet: Graph attention based point neural network for exploiting local feature of point cloud," arXiv preprint arXiv:1905.08705, 2019.

[3] H. You, Y. Feng, R. Ji, and Y. Gao, "Pvnet: A joint convolutional network of point cloud and multi-view for 3d shape recognition," in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1310-1318.

Acknowledgment

The project sponsored by the National Key Research and Development Program (No.2016YFB0502002) and the Fundamental Research Funds for the Central Universities under Grant 2019PTB-010.

The visualization results of the soft attention fusion block


airplane1
0.4692
0.1915
0.1719
0.2728
0.3144
0.1332
0.0462
0.1557
0.2445
0.3538
0.2637
0.1849
airplane2
0.6290
0.2185
0.1214
0.2560
0.2063
0.0602
0.0817
0.0539
0.2827
0.2859
0.3145
0.1178
car1
0.1336
0.7316
0.2977
0.1144
0.2885
0.5806
0.0960
0.4188
0.3698
0.1510
0.4149
0.7393
car2
0.2873
0.5284
0.6783
0.2650
0.6790
0.6826
0.2657
0.5051
0.7147
0.2152
0.6917
0.5384
monitor1
0.1493
0.6429
0.6769
0.5726
0.6995
0.6171
0.2289
0.6040
0.7360
0.7087
0.7323
0.5481
monitor2
0.1513
0.5348
0.7424
0.6232
0.7891
0.6171
0.2858
0.6191
0.7038
0.7103
0.6778
0.5421