An Improved Bilinear Pooling Method for Image-Based Action Recognition

Wei Wu; Jiale Yu
Inner Mongolia University

Abstract

Action recognition in still images is a challenging task because of the complexity of human motions and the variance of background in the same action category. And some actions typically occur in fine-grained categories, with little visual differences between these categories. So extracting discriminative features or modeling various semantic parts is essential for image-based action recognition. Many methods apply expensive manual annotations to learn discriminative parts information for action recognition, which may severely hinder potential applications in real life. In recent years, bilinear pooling method has shown its effectiveness for image classification due to its learning distinctive features automatically. Inspired by this model, in this paper, an improved bilinear pooling method is proposed for image-based action recognition. The previous bilinear pooling approaches contain lots of noisy background or harmful feature information, which limit their application for action recognition. In our method, the attention mechanism is introduced into hierarchical bilinear pooling framework with mask aggregation. The proposed model can generate the distinctive and RoI-aware feature information by combining multiple attention mask maps from the channel and spatial-wise attention features. Specifically, our method makes the network to pay more attention to discriminative region of the vital objects in an image. We verify our model on the two challenging datasets:1)Stanford 40 action dataset and 2) our action dataset that includes 60 categories. Experimental results demonstrate the effectiveness of our approach, which is superior to the traditional and state-of-the-art methods.

Methods

The model is divided into two important components for the detailed introduction. Firstly, the channel and spatial-wise attention mechanism is applied to enhance the feature representation ability. The Fig. 2 show the channel-wise and spatial-wise attention module respectively. Secondly, an aggregated mask is generated by integrating multiple attention feature maps to extract the robust RoIs which reduce the noisy background. The overall architecture of the model is shown in Fig. 1.

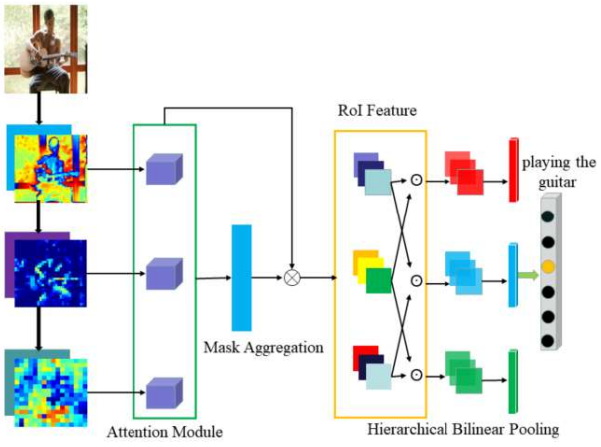

Fig. 1 The architecture of the improved bilinear pooling model.

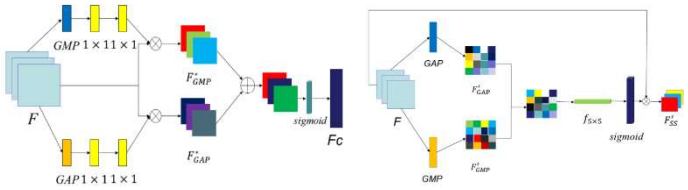

Fig. 2 Channel-wise and Spatial-wise Attention Module.

Results

1、we establish a new image-based action dataset containing 18046 images and 60 action classes.
2、We achieved the best performance on the Stanford 40 actions and our 60 custom action dataset. Fig. 3 shows that the proposed method gains the best results of 85.24% in the Stanford 40 Actions dataset against all previous methods. We also evaluate the model on our 60 action dataset. Our model obtained 84.57% mAP on this dataset.
3、We also conduct the supplementary experiments about mask aggregation on the Stanford 40 action dataset. Fig. 4 visualizes image masks generated on some samples in the Stanford 40 action dataset. It implies that mask aggregation effectively reduces the impact of the harmful background information and generate a robust RoI feature to improve the accuracy of action recognition.

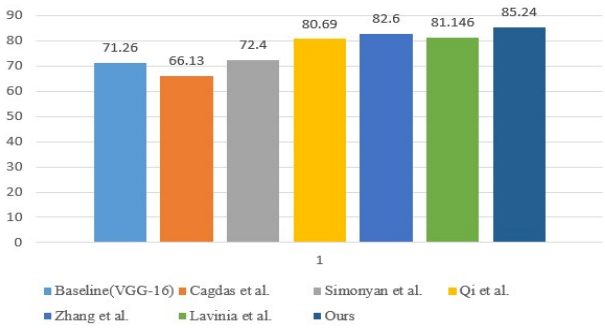

Fig. 3 Performance (mAP) on the Stanford 40 Actions dataset.

Fig. 4 Visualization of the aggregated mask on some samples in the Stanford 40 action dataset.

Conclusions

1、An Improved Bilinear Pooling with Mask Aggregation method can extract more distinctive features through the channel and spatial-wise attention module and uses a multi-level mask aggregation which generates a RoI-aware feature to restrain background interference.
2、Extensive experiments are evaluated on two datasets, ranging from 40 to 60 action categories, which validate the effectiveness of our approach. Experiments on large-scale dataset further demonstrate that our model has good generalization ability.

Contact

Wei Wu
Department: Inner Mongolia University
Email: cswuwei@imu.edu.cn

References

1. Yu, Chaojian, et al., "Hierarchical bilinear pooling for fine-grained visual recognition," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, doi: https://doi.org/10.1007/978-3-030-01270-0_35.
2. G. Gkioxari, R. Girshick, P. Dollár and K. He, " Detecting and Recognizing Human-Object Interactions," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 8359-8367, doi: 10.1109/CVPR.2018.00872.
3. K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proceedings of the International Conference on Learning Representations, 2015.
4. Woo, Sanghyun, et al., "Cbam: Convolutional block attention module," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, doi: https://doi.org/10.1007/978-3-030-01234-2_1.
5. B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas and L. Fei- Fei, "Human action recognition by learning bases of action attributes and parts," 2011 International Conference on Computer Vision, Barcelona, 2011, pp. 1331-1338, doi: 10.1109/ICCV.2011.6126386.
6. Wenqian Wang， Jun Zhang andFenglei Wang, " Attention Bilinear Pooling for Fine-Grained Classification," in Symmetry, vol. 11(8), pp. 1033, 2019, doi: https://doi.org/10.3390/sym11081033.