Non-local Neural Networks

Liangjie Cao

Aug. 13, 2018

1. Introduction

It opens up a new direction to solve the long-distance dependence in space-time domain in video processing. In this paper, non-local averaging method is used to deal with the relationship between local features and feature points of panorama. This kind of non-local operation can be easily embedded into the existing model, and has achieved good results in the video classification task, and surpassed the Master-CNN of his ICCV best paper in the static image recognition task. And surpass CNN to overcome the shortcomings of CNN network being too concerned about local features. Inspired by the application of NL-Means in image denoising, the task of serialization is to consider all feature points for weighted computation, which overcomes the shortcoming of CNN network that pays too much attention to local features.

2. Motivation

The word ”Non-local” is literally translated into ”non local”. Personal understanding is that when feature extraction, the current input data feature calculation should consider the information of other input data. For example, the focus of non-local operations is how to establish the connection between two pixels with a certain distance on the image, how to establish the connection between the two frames in the video[3], how to establish the connection between different words in a paragraph.

A typical CNN network is accumulated by a series of convolution operations. For CNNs using images, each convolution operation captures only local information of the input data. The whole network obtains a wider range of information extraction through the gradual accumulation of local operations. RNN processes the sequence input (such as the time series of each frame of video or the spatial sequence of a column of pixels on the picture) in a circular manner, thus fusing non-local information. This paper presents three disadvantages of CNN and RNN in the fusion of non-local information: 1. inefficient computation; 2. more difficult optimization; 3. non-local features of information transmission is not flexible enough, not powerful enough. Of course, this is also because the original intention of CNN and RNN is not to integrate non local information.

In this paper, the author proposes a network architecture of non-local block (NL block) to help the deep network better integrate non-local information. This is very important for some problems.

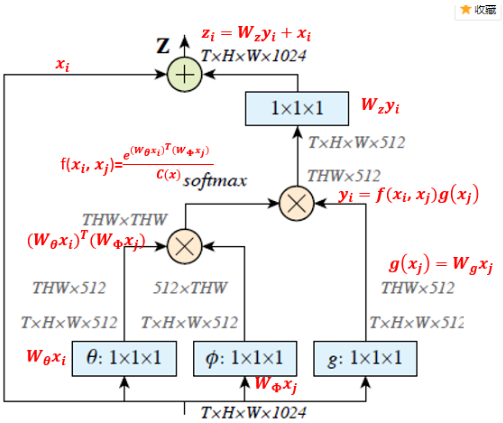

Figure 1. Model of non-local

3. Experiment

Experiments are carried out on video classification, object detection and object instance segmentation tasks that require non-local information association. The ablation study was conducted on the Kinetics dataset to examine the effectiveness of the details of the NL block. The results will not be repeated. [2]

There are different definitions of F (.) in NL blocks, but for better visualization use embedded Gaussian + dot product, the method shown in the formula mentioned above.

The position of NL block is placed in the backbone of the network: put it in the shallow layer, and increase in the upper reaches.The role of NL block deepening: for the shallow backbone network, deepening NL block can improve performance. It is difficult to improve performance for larger and deeper networks, either by adding NL blocks or by deepening the depth of the backbone network.(video task) NL block is better than time alone in time domain or space domain.(video task) compared with C3D [1]: faster and better than C3D. We presented a new class of neural networks which capture long-range dependencies via non-local operations. Our non-local blocks can be combined with any existing architectures. We show the significance of non-local modeling for the tasks of video classification, object detection and segmentation, and pose estimation. On all tasks, a simple addition of non-local blocks provides solid improvement over baselines. They hope non-local layers will become an essential component of future network architectures.

References

[1] S. Karaman, L. Seidenari, and A. Del Bimbo. Fast saliency based pooling of fisher encoded dense trajectories. In *ECCV*, 2014. 2

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[3] L. Wang, Y. Qiao, X. Tang, and L. Van Gool. Actionness estimation using hybrid fully convolutional networks. In *CVPR*, 2016. 1