Student ID : 20194293

Name : Go, Kyeong Ryeol

**[AI 502] Group Normalization**

1. Paper Summary

Batch Normalization has resolved the 'Internal Covariate Shift' problem by adding a normalization layer right before the activations for whitening. This enables to train various deep neural networks and now becomes a milestone technique in many papers. However, since the mean and variance are estimated in a mini-batch unit, the training gets unstable as the batch size gets smaller. This blocks the use of Batch Normalization in a higher capacity network as a small batch size is inevitable due to the memory consumption. Many computer vision tasks like detection, segmentation and video are related with this issue. Furthermore, there is no normalization performed in the test where the statistics is pre-computed by the all the training examples and this may occur the inconsistency problem when the target data distribution is somewhat different. Therefore, the author presents Group Normalization as an alternative which calculates the estimate of mean and variance in a group unit.

For analysis, let's assume 2D images that consists of 4 feature axes which are N(batch), C(channel), H(spatial height) and W(spatial width). Then, Batch Normalization computes the mean and variance along the (N, H, W) axes. Several related works proposed other normalization methods along different axes such as Layer Normalization(C, H, W) and Instance Normalization(H, W). Group Normalization, which made a nice consensus of these two, defines a group by dividing the channels and normalize the features within the group. As an analogy, the mean and variance are computed along the (G, H, W) where G is kind of sub-axis of C. Group Normalization is more flexible than Layer Normalization in the sense that it can learn different distribution of different groups within the same batch. Also, it can exploit the dependence among channels All methods mentioned above learn a per-channel linear transform to compensate for the possible cost of representational ability by introducing the additional parameters, beta and gamma.

The experiment in ImageNet classification task with ResNet-50 shows that even if the validation error of Batch Normalization is 0.5% smaller with the moderate batch size, Group Normalization behaves very stably over a wide range of batch sizes. Moreover, the author evaluated the error rate by differing the number of group size. This was further shown in object detection and segmentation task in Microsoft COCO, C4/FPN backbone datasets and in video classification in Kinetics dataset.

2. Discussion

Here, I want to offer two discussion points. To begin with, what if Batch Normalization and Group Normalization are used in a network in alternating manner? The experiment shows that none of the two are dominant. As the ensemble method implies, better performance may be achieved. Specifically, I set the hypothesis that Batch Normalization may be effective in fully connected layers while Group Normalization would be better in convolutional layers. Next, what if a group is defined by dividing not only the channels and but also the batches? Just as Group Normalization is derived by combining Layer Normalization and Instance Normalization, another nice normalization method may be created in this way.