Norm matters: efficient and accurate normalization schemes in deep networks

Abraham Jose
ID :5068109, CAP6614
abraham@knights.ucf.edu

1 Summary

Batch-normalization is a useful technique to converge the model soon and for superior performance. However, effects of batch-normalization in a network hasn't been fully studied yet. Here, author presents a novel view on the purpose and function of normalization methods and weight-decay, as tools to decouple weights' norm from the underlying optimized objective. They suggest alternatives to L2 batch-norm, using normalization in L1 and L spaces that can substantially improve stability and low-precision implementation has computational and memory benefits as well.The technique improves the performance in large-scale tasks.

2 Strengths of the proposal

1. Empirically proves that batch normalization is scale-invariant, requires high precision and task oriented, some of the findings which are not studied earlier. It is important to understand why the batch-normalization works to improve the technique further.
2. Summarize that by using batch-normalization and weight decay, we can fix the norm in small range so that we can take more stable steps but requires high precision and numerically not stable. For low precision operation, they device a technique for using L1 to L with improved stability and performance.
3. Proves that a bounded weight normalization method will always generalize better with comparable performance with batch normalization. This will be well suited for improved learning tasks such as reinforcement-learning and temporal modeling.

3 Weaknesses of the proposal

1. The results are mostly empirical and based on observations rather than strong proofs that is required to support the claim.
2. Author identifies batch-normalization as a task oriented optimization technique. The proposed bounded weight normalization method should also be suffering from the same problem of not able to generalize enough. There should be enough support to this claim.

4 Results

The author was able to experimentally prove that it is possible to use an alternative to batch-normalization with low precision and which are numerically stable while training. Also, they were able to train successfully train 8-bit Resnet-18 and 16-bit Resnet-50 models without any drop in accuracy. They also proved that the L2 can be mimicked by multiplying with a constant $\sqrt{\pi}$, which will shows same behaviour as L2 norm which is an interesting result.

5 Discussion

The optimization and the data-set takes a huge toll while training a deep neural network model. What all are the required features that we are looking at in a optimization technique? How much the norm and the techniques like batch-normalization and weight decay helps the model to learn and do we require them even if we could find an absolute perfect optimization technique?

CAP 6614: Current Topics in Machine Learning (Fall 2019), UCF.