Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

Authors: Sergey Ioffe, Christian Szegedy

Publisher: Computing Research Repository
Year: 2015

Keywords: Batch Normalization, Inception, Covariate shift

1. Abstract:

“Training Deep Neural Networks is complicated by the fact that the distribution of each layer’s inputs changes during training, as the parameters of the previous layers change. This slows down the training by requiring lower learning rates and careful parameter initialization, and makes it notoriously hard to train models with saturating nonlinearities. We refer to this phenomenon as internal covariate shift, and address the problem by normalizing layer inputs. Our method draws its strength from making normalization a part of the model architecture and performing the normalization for each training mini-batch. Batch Normalization allows us to use much higher learning rates and be less careful about initialization. It also acts as a regularizer, in some cases eliminating the need for Dropout. Applied to a state-of-the-art image classification model, Batch Normalization achieves the same accuracy with 14 times fewer training steps, and beats the original model by a significant margin. Using an ensemble of batch- normalized networks, we improve upon the best published result on ImageNet classification: reaching 4.9% top-5 validation error (and 4.8% test error), exceeding the accuracy of human raters.”

2. Introduction:

- The difficulty of training for deep neural networks had been so relevant in many of the fields, it is applied. To overcome that obstacle, there had been significant approaches which increased the efficiency of training, in which this article had also shared one of them.
- The fact that the parameters of the previous layers change continuously updating would result in the lower learning rates and eventually affect the training time.
- So, the paper suggests that the normalization on the mini batch would gradually increase the learning rate of training. It is also said that the batch normalization resulted in 14 times lower steps of training .

3. Related Works:

- The change in the input distribution due to the change in learning system is referred as the covariate shift by Shimodaira. The approach had been focused on reducing the covariate shift.
- Whitening is the approach of transforming to result in the mean of zero and variances are 1 by LeCun et al., had also been adapted by the authors for the approach.
- To overcome overfitting of the model, the dropouts had been one of the methods in deep learning. In this paper, the authors had endeavoured to eliminate dropouts which would result in lower strength of the networks.
- One of the relevant approaches is the Inception architecture by Szegedy et al., which consists of simple CNN architecture which had shown very good results in ILSVRC. The work had been done in this article to extend it to Batch Normalization to emerge into more accurate results.

4. Content and Approach:

- To minimize the loss functions we use Stochastic Gradient Descent and AdaGrad algorithms in the state of the art.
- The mini batches were used to estimate the loss, by gradient functions on the training sets. The other advantage with the minibatch is that the computations will be decreased compared to compute each input example.
- The covariance shift is referred to the continuous upgrade of the previous layers due to the new distribution. This covariance shift had to be decreased to increase the learning rate.
- In implementation, the vanishing gradients and the saturation problems were referred by Rectified Linear Units(ReLU).
- The change of distributions in the deep network internal nodes were referred as the internal covariate shift. The new mechanism Batch normalization(BN) introduced by the authors would decrease the internal covariate shift and increase the training rate.
- BN is also advantageous in dealing with the gradient flow in the network, which increases the independence of gradients of the initial values, which decreases the possibility of divergence.
- The approach of whitening is carried out in order to decrease the internal covariate shift. Whitening is the process of transforming the vectors to means and variances of 0 and 1 respectively.
- The whitening can be done at each step of training or particular intervals. Some calculations were done to represent that the change in normalizations would not affect the loss of the output layers.
- If the norms were considered for the training, then the covariance was calculated and the inverse square root would result in the whitened activations.

- The whitening process for every layer is also expensive and differentiatives were not possible for all types of layers. For that cause, two simplifications were done:
  - Instead of whitening the input and output features jointly, each scalar feature was normalized independently resulting in the zero mean and unit variance.
  - As the stochastic gradient descent was adapted for the mini batches, the mini batches would estimate the variance and the mean of every activation.
- The whitening is done on the mini batches which is referred as the Batch Normalizing Transformation. The means, variances, scales and shifts were calculated for each batch considered.
- The normalized activation distribution values would have the zero mean and unit variances until or unless they had selected from the same batch.
- In the process of backpropagation, the gradients had to be calculated, relative to the parameters for BN transform.
- This makes the change inputs from x to BN(x). So, now the training is carried out on the BN(x) which would accelerate the training.
- The experiments were done on the ImageNet dataset with Inception network. Moreover just to implement the Batch Normalization would not result in the better results. In addition, they had changed the following training parameters.
  - Increased the learning rate
  - Eliminating Dropouts
  - Cutting down the L_2 weight standardization
  - Quickening the learning rate decay
  - Rearranging the training examples frequently
- The training was also done on different networks such as Inception, BN-x5, BN-x30, BN-x5-Sigmoid etc.,

5. Results:

- As the purpose to be solved, the BN had lessen the internal covariate shift and in fact made more stabilized distribution with sigmoid function.
- The results were based on the least steps taken by the architecture to attain the highest accuracy of 72.2% on LSVRC2012 data.
- BN-x5 had taken the least steps of 2.1 x 10^6 to attain the accuracy of 72.2%, where as the previous state of the art Inception architecture had taken around 31 x 10^6 steps.
- In the top-5 error reports too, the model had resulted in 4.82% which had been lower than the previous best results.

6. Conclusion and Future works:

- It can be observed that the batch Normalization method had been performed better with the previous state of the art Inception architecture on the ImageNet dataset.
- It is also obvious that implementation of the model for pretraining has been resulting in the significant outputs.
- Furthermore, we can notice the stable distribution of the sigmoid functions during training with the applications of Batch Normalization for activation values.
- In future, the extended works can be focused on implementation of Batch Normalization on Recurring neural networks which would result in exhibiting the severity of reducing the internal covariate shift and boosts the gradient propagation.