Data Set Normalization for High Accuracy

Andarta Fardhanul Khoir[1],

[1] The Australian National University, u6297117@anu.edu.au

Abstract. Using normalized data set to train a Convolutional Neural Network (CNN) model will give positive impact. Regarding normalization, the main objective of this paper is to ensure that normalization gives positive impact in loss reduction during training and high accuracy during test. To reach this objective, normalized and not-normalized data sets are used to train CNN models. During training, models trained using not-normalized data shows significant loss reduction. However, using normalized data in building CNN model is still better option since the test result shows that it gives best test result. Therefore, it can be said that although using data normalization for training does not always give better loss reduction during CNN training, its test result is better.

Keywords: Convolutional Neural Network (CNN), normalization, loss, accuracy, training, test.

1. Introduction

Data set used for neural network training must be reasonable enough to be used. At this case, reasonable means that the model can have high test accuracy without being trained in many epochs. One method that can be applied to make a data set reasonable is normalization. In fact, Bustos and Gedeon (1995) has tried this normalization for altitude data concerning the logistic function. Moreover, Nicholas (2012) also utilize data set normalization to make training of neural network faster.

Regarding normalization, this paper, indeed, tries to ensure if normalization of data set gives positive impact during CNN training and test. In accordance to this purpose, firstly, methodologies will be briefly described. At this section are the description of data set normalization, CNN topology, training setup, and the investigation that will be done. At next section, results and discussion will be explained. Lastly, the importance of data normalization will be concluded.

2. Methodologies

The main objective of this paper is to answer if normalized data set gives significant impact during neural networks training, especially for the changes of weights of hidden neurons. Thus, to make sure that this question has an appropriate answer at the end, some conditions should be made. Firstly, appropriate structure of neural network should be made. Next, data set, which is the main topic in this paper, will be given some treatments. Lastly, investigation carried in this paper will be briefly described.

2.1. Data Set Normalization

Data set used in this paper is Caltech101 (L. Fei-Fei, Fergus, & Perona, 2004). Originally, this data set consists of 101 classes. Each class has about 40 to 800 images. However, in this paper, only 8 classes are used: pagoda, panda, pigeon, pizza, platypus, pyramid, rooster, and snoopy. Total images from all classes is 358. The reason of why not all classes are used is the fact that the purpose of this paper is only to ensure if normalization gives positive impact.

This data set are then randomly selected to be used for training, validation, and test. 60% data set is used for training. The rest is equally split into validation and test. Surely, all classes are included in training, validation, and test.

As mentioned before, the main purpose of this paper is to make sure if normalization gives positive impact. Therefore, this data set will be normalized to train some models. Normalization implemented to the dataset uses equation below.

$$X' = \frac{X - min(X)}{max(X) - min(X)} \qquad \textbf{(1)}$$

X in that equation represents values in same attribute. Using equation above, values in range from 0 to 1 will be achieved. Indeed, for current data set, this equation is applied to all pixels in each channel in each image.

2.2. CNN Topology

In this paper, CNN model consists of 2 convolutional layers with 5 kernels. Since the images used as input is an RGB image, first convolutional layer has 3 input channels. For each convolutional layer, activation function utilized is ReLU. Further, method used in the pooling layer is max pooling. At last, there are fully connected layers consist of three linear layers with ReLU as activation function. In this topology, sigmoid function is not used because sigmoid activation function cannot handle many hidden layers. If many hidden layers are used, sigmoid activation function may easily cause vanishing gradient (Glorot & Bengio, 2010).

2.3. Training Setup

To make sure if normalization absolutely gives significant impact, comparison is needed. Thus, two data sets are used. The first data set is the data set containing not-normalized images. At the other hand, second data set is the normalized data set. However, although two data sets are used, both data set have same instances.

During training, two learning rate are used. These learning rate are 0.0005 and 0.005. Surely, at each learning rate, two models are trained based on data sets used.

To make sure that models achieved are the best model, model is trained 5 times using training data set. After training, each model is validated using validation data set. The best model is the model that has highest accuracy in predicting validation data set. At last, four models are achieved. Surely, they have different combination of learning rate and data set.

2.3. Carried Investigation

Two parameters will be analyzed for investigation. First parameter is loss during training. Surely, loss should be decreased regularly during training. Second parameter is accuracy of test result. Of course, test is only performed once after the best model is picked.

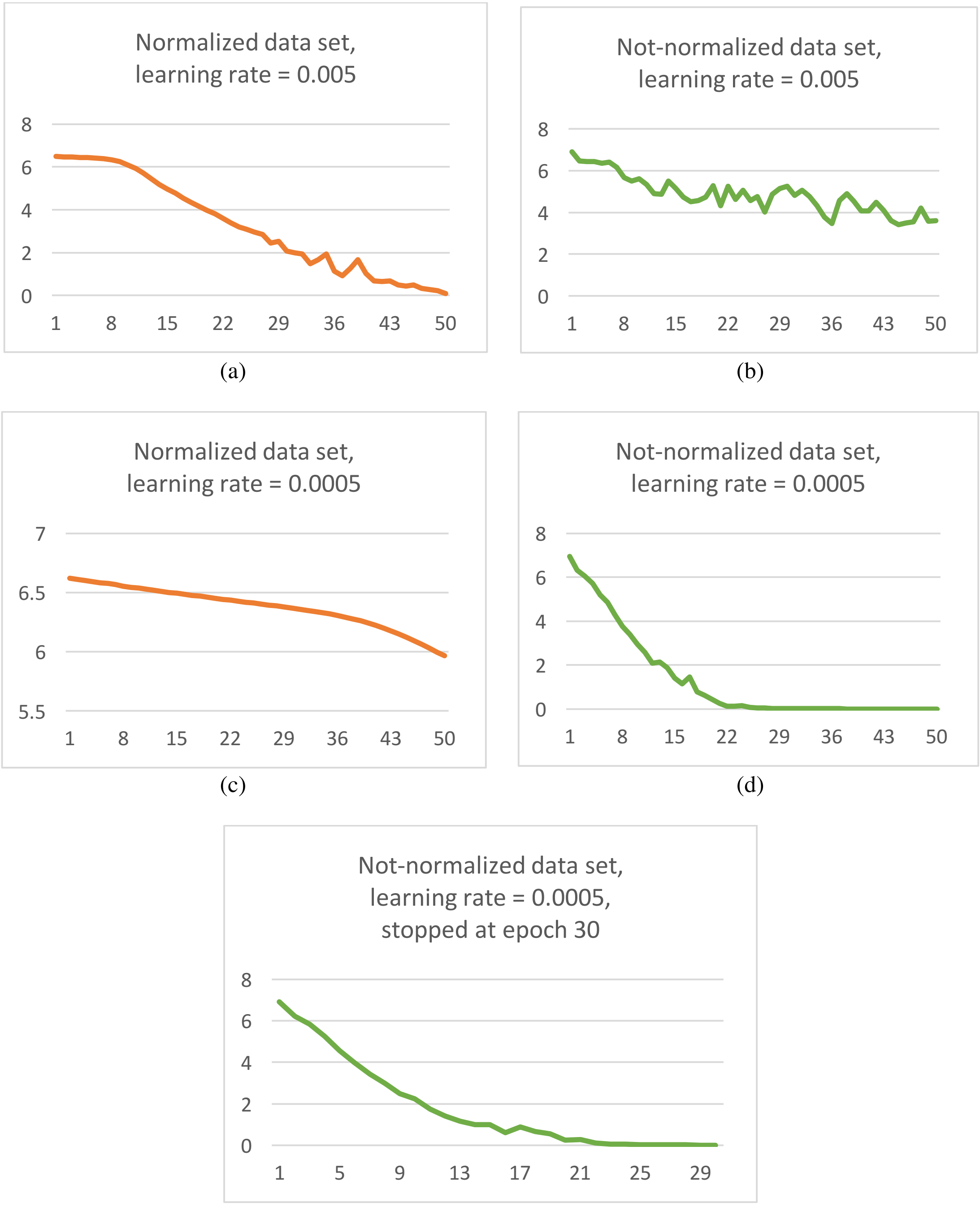
3. Results and Discussion

3.1. Training

Figure 1(a) shows loss during training using normalized data set. This training is performed in 50 epochs. During this training, loss reaches 0.1 at the last epoch. At the other hand, using not-normalized data set (see Figure 1(b)), loss can only be reduced until 3.6. Based on this fact, it can be said that using normalized data is the best option for training CNN.

Contrastly, Figure 1(c) shows different result. In this figure, using normalized data set does not show significant loss reduction. At epoch 50, loss is only reduced until 5.97. However, the use of not-normalized data has given better loss reduction (see Figure1(d)). In fact, it seems like the model may do overfitting during training. Therefore, particularly at learning rate 0.0005, training of the CNN model using not-normalized data is stopped at epoch 30 (Figure 1(e)). This training is stopped at epoch 30 because it has reached very low loss, 0.1. Probably, if the training is continued, the CNN model will be overfitting.

So far, during training, it can be said that the model trained using not-normalized data with learning rate 0.0005 may potentially give best result in test because it has the best loss reduction during training.



(a)

(b)

(c)

(d)

(e)

Figure 1. Loss during training:
(a) using normalized data set with learning rate 0.005;
(b) using not-normalized data set with learning rate 0.005;
(c) using normalized data set with learning rate 0.0005;
(d) using not-normalized data set with learning rate 0.0005;
(e) using not-normalized data set with learning rate 0.0005, but stopped at epoch 30 to avoid overfitting.

3.2. Test

It is interesting to compare what all models gain during test with what all models gain during training. Although CNN model trained using not-normalized data gives best loss reduction at learning rate 0.0005, it does not give the best result during test. It only achieves the accuracy 62% during test. However, CNN model trained using normalized data at learning rate 0.005 gives the best result, 64%.

Table 1. Accuracy of test result of two neural networks

Dataset	Learning Rate	Accuracy of test result
Normalized	0.0005	23%
	0.005	64%
Not-normalized	0.0005	62%
	0.005	34%

Based on this result, it can be said that using normalized data during training is the best option in building CNN model. However, result achieved in this paper is still low. Ajeesh, Indu, and Sherly (2014), although using different model, has successfully gained accuracy until 79.2%.

4. Conclusion and Future Work

It can be concluded that data set normalization should be performed before training a neural network. At certain learning rate, using normalized data set can give acceptable loss reduction and high test accuracy. Although in this paper the loss reduction during training may not be the best, the result of test still shows that using normalized data is must.

Surely, what author does in this paper has many shortcomings. One of these is the fact that the number of instances used is too small. Thus, using CNN model created by the author, the result gained if all Caltech101 images are used is still unknown.

5. References

Ajeesh, S. S., Indu, M. S., & Sherly, E. (2014, 7-8 Feb. 2014). *Performance analysis of classification algorithms applied to Caltech101 image database*. Paper presented at the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT).

Bustos, R. A., & Gedeon, T. D. (1995). Decrypting Neural Network Data: A GIS Case Study *Artificial Neural Nets and Genetic Algorithms* (pp. 231-234). Vienna: Springer.

Glorot, X., & Bengio, Y. (2010). *Understanding the difficulty of training deep feedforward neural networks*. http://proceedings.mlr.press/v9/glorot10a.html

L. Fei-Fei, Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *IEEE, CVPR 2004*(Workshop on Generative-Model Based Vision).

Nicholas, N. K. R. (2012). *Forecasting of Wind Speeds and Directions with Artificial Neural Networks*. (Master), Lappeenranta University of Technology, Lappeenranta, Finlandia. Retrieved from https://www.doria.fi/bitstream/handle/10024/98414/Rotich %28M. Sc Thesis%29.pdf?sequence=2