EasyChair Preprint

№ 576

Escape the local minima by error reflection

Liyao Gao

October 13, 2018

Escape the local minima by error reflection

Liyao Gao
Purdue University
gao463@purdue.edu

Abstract

For the current deep learning, one of the most important questions is: how to make neural network escape from local minima or saddle points. People tend to believe that a local minima is already able to reach satisfaction. In this paper, we provided theoretical analysis for the situation when the neural network is trapped into local minima. We would try to criticize the point that "local minima is good enough." Furthermore, based on the property, information forgetting in local minima, that we investigated in this paper, we provide a possible method to solve this problem with error reflection. Our experiments provide strong evidence of this method where it can lower the loss by 99% for our designed function approximation tasks. It can serve as evidence of our theory in two-dimensional space. Our testing result on image recognition task also shows its superiority that can reach 98.81% in Fashion-MNIST datasets with parameter size of 3000.

1 Introduction

Many deep neural networks can reach a satisfying performance in many tasks, including image recognition, speech recognition, and machine translation [1]. A huge point that hinders the further improvement of neural networks is the local minima and the saddle points. Clearly, we hope to allow the network to find the optimal parameter. This paper will try to provide a further investigation into the situation when the network is trapped into local minima, and discuss a possible method to escape from it.

Researchers have been focused a lot in this field. It is impossible to find the global minima since it is NP-hard. The massive amount of parameters in neural networks make it even worse to find the optimal by testing all the discrete combinations. Also, optimization methods such as SGD is easy to enter local minima [2]. Other optimization methods, such as Newton's method, are hard in computation. Based on the above limitations, current training methods is hard to escape from local minima once it passes nearby.

In this paper, instead of merely recognizing the local minima as the point when the gradient is zero, we try to view it from the aspect using the training data. We suppose another definition of local minima that only fixed part of the data in training set are fitted near local minima. In other words, only part of the pattern of the training set is learned by the network, and there exist typical errors in the region of local minima. The summarize the reason as active information forgetting for the neural network. The effect of error cases which should originally change the network's parameter is counteracted by the cases which already well-fitted in the current model. In other words, the neural network will actively forget the information of the error cases.

To solve the situation above, we provide a possible solution by reflecting the errors. Error reflection is an important learning technique for human [3]. It is simple but powerful in our situation. Based on our experiments, in function approximation task, it can lower the loss by 90% compared to the original network. Also, in image recognition, we can reach the test accuracy of 95.88% for Fashion-MNIST, which is comparable to the performance of ResNet and VGG. A point that we should note here is our

Preprint. Work in progress.

model has much lower parameter size, only 3000 parameters, with an easier training process. Due to the page limit, the experiment in Fashion-MNIST is not fully extended in this paper.

2 Intuition

We provide a simple case with local minima with ReLU function as the intuition of our work. This part can explain another view of information forgetting in neural network training. The example will be described as follows.

In our experiment, suppose we have a function approximation task with a revised ReLU function. Consider the case $f(x) = \lfloor max\{x - b, 0\} \rfloor$. $f(x)$ is the floor, which defined by $\lfloor f(x) \rfloor = max\{n \in Z | n \leq f(x)\}$, of the output of a simplified ReLU function with only one parameter b. Our training data is generated by the following piecewise function that its domain is $D = \{x | x \leq 1 \, or \, x = 4 \, or \, x = 4.5\}$. The function which generates the data is:

$$\begin{cases} 0 & x \leq 0 \\ 1 & x = 1, \, x = 4, \, x = 4.5 \end{cases}$$

, where is represented in figure 1.

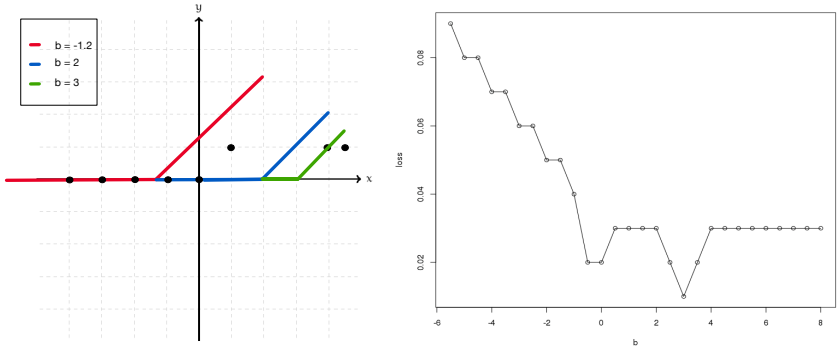

Figure 1: Left: Approximation function with different b with training data. Right: Loss function plotting with local minima and global minima

As we can see in Figure 1 (Left), there exist a local minima around $b = 0$ and global minima at $b = 3$. When the parameter is close to local minima, anywhere else with certain length has a worse performance compared to the local minima. It is a huge trouble in the phase of optimization.

A simple method to use in this situation is to train a new network to learn the pattern based on the forgot information before. Consider the example we use in intuition, the network will reach 100% accuracy with the following solution even if the first network is trapped into local minima. It can be obtained by the method of error reflection which we will discuss later:

$$f(x) = \begin{cases} \lfloor max\{x, 0\} \rfloor & x \leq 1 \\ \lfloor max\{x - 4, 0\} \rfloor & 1 \leq x \end{cases}$$

3 Information Forgetting in Local Minima

In this section, we would like to describe a general situation in local minima. The network will actively forget the error cases which caused by other data already fit in the current pattern provided by the network.

According to the gradient descent that $\theta_{n+1} = \theta + \eta \nabla L(\theta)$. The gradient, $\nabla L(\theta)$, is small when approaching local minima. Based on current techniques, it is hard to escape local minima. In other words the training will be ended when close to local minima. However, it is clear that we hope to reach a higher performance.

In fact, the local minima is a huge cause of information forgetting in the training process. In a training process, every information that carried by the training data is going to be stored into the trained models by changing its parameter. However, consider the situation when $b \in (-1, 0)$, in this local minima, in fact, the model will continuously make wrong predictions for the points $x = 4$ and $x = 4.5$. Once the optimization process is closed to this local minima, and, in the end, trapped into this local minima. The training process is repeated the following process.

1. Wrong predictions will try to update θ out of it by the step of $\eta \nabla L(\theta)$.

2. Under this condition, after the update in parameter in step (1), $\eta \nabla L(\theta)$ is still pointing to the local minima. Note that in this situation that the network is actively forgetting the information of errors.

We can conclude that when an optimization process of a neural network is trapped into local minima, the errors will be actively forgotten by the network. It only fits a specific part of the trend in the training set. The reason for this active forgot caused by the data already in following the current trend of the model. No matter how more training process is performed, the information will keep losing in this situation. This could serve as another view of local minima.

This phenomenon can cause the limitation in performance. The local minima tend to learn regular pattern in the task while there can be more than one patterns in a learning task. This could serve for the reason that neural networks will have low performance in the large-scale learning task.

4 Error Reflection & Experiment

4.1 Method of reflecting the errors

A general neural network will be initially trained for a learning task. Then, the error cases of the general network will be collected. The algorithm will classify the error cases into different clusters and initialize a corresponding number of neural networks to be trained by the error clusters. After the above process, a classifier will be trained to determine which network to use for the incoming data. In our implementation, we use ISODATA, The Iterative Self-Organizing Data Analysis Technique, to further split the error cases into different clusters. Figure 2 can show the architecture of our method.

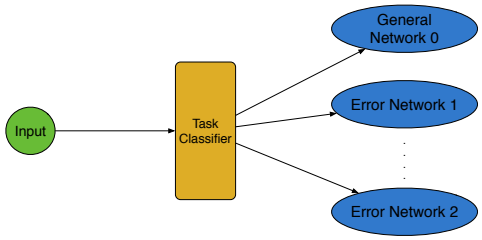

Figure 2: Architecture of error reflection

4.2 Approximation of function in two dimensional space experiment

In this experiment, we are trying to test the ability of error reflection based on a task of approximation of function in two-dimensional space. The base neural network we choose here is a feed forward network. We implemented the network and make error reflection approximate the assigned functions based on learning data. We tested the result on linear continuous function, higher-order continuous function, linear discontinuous function and higher-order discontinuous function. The base network we choose here is a simple Feed Forward Network with one hidden layer using RELU as activation function. The Adam is used here as the optimizer with the learning rate of 0.0001. The discontinuous function is trying to simulate the situation that different tasks are involved in the training data.

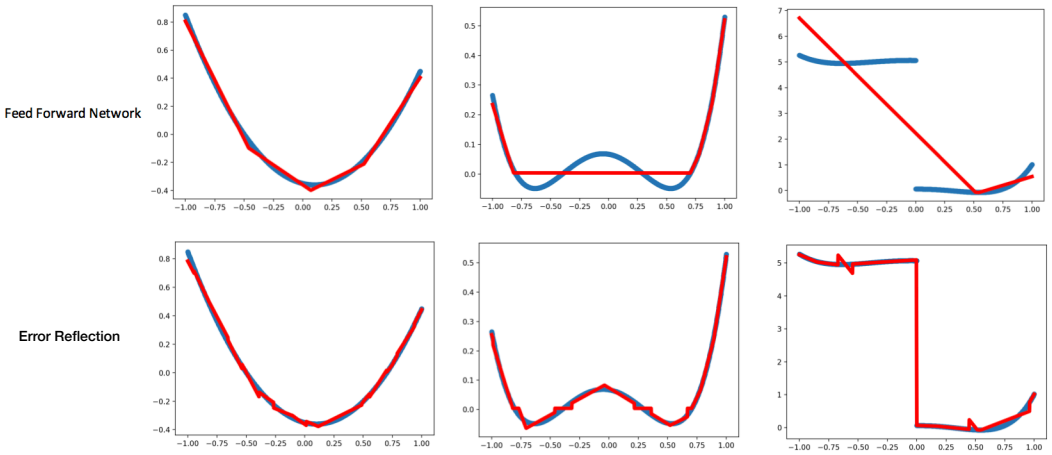

Figure 3: Performance of Single Feed Forward Network on approximation task between before and after applying error reflection

Preformance of Error Reflection and SingleNN			
Function	BaseNN Loss	Reflection Loss	Loss reduce %
$f(x) = x$	3×10^{-9}	3×10^{-10}	90%
$f(x) = x^3 - 0.2x - 0.35$	4×10^{-4}	2×10^{-4}	92%
$f(x) = x^4 + 0.2x^3 - 0.67x^2$	1×10^{-3}	5×10^{-5}	95%
$f(x) = x^4 + x^3 - 0.6x^2, x \in [-1, 0)$ $f(x) = x^5 + x^4 - 0.5x^3, x \in [0, 1)$	1.4	1.4×10^{-2}	99%

Based on the result, we can find out that the error reflection is able to have a better performance compared to the base neural network (SingleNN). An interesting observation here is: as the function is more complex, the base network became worse in performance. However, the reduction in loss for error reflection is becoming higher. Especially in the case of the discontinuous function, it can be approximated with a much higher performance. This experiment show explains how error reflection works (see Figure 3) and provide strong evidence on the situation that the error reflection can have a much better performance as the function is becoming more complex.

5 Conclusion

The phenomenon of information forgetting influences the performance of local minima from the view of the training set. The information of error cases will continuously be ignored in the training process after entering the local minima. Methods to retrain the error cases do help to increase the performance significantly. The loss reduced is higher than 90% for function approximation, and the performance can reach 98.81% for Fashion-MNIST.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[2] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

[3] Liyao Gao. Cortex neural network: learning with neural network groups. *arXiv preprint arXiv:1804.03313*, 2018.