Identification of the Challenges of Local Minima in Recurrent Neural Networks

Ochonogor Donpaul[1], Elebra Charity[2] and Friday E.Onuodu[3]

[1]*Department of Computer Science, School of Postgraduate Studies, Ignatius Ajuru University of Education Rivers State, Nigeria*

[2&3]*Department of Computer Science, University of Port Harcourt, Rivers State, Nigeria*

Abstract: **- Numerous researchers have recently based efforts on the weight of repeated neutral networks through the creation of efficient algorithms, primarily to optimized schemes. For feed forward network, the learning algorithm can become stuck in local minima during gradient descent. This research focuses on recurrent neural networks, local minima in neural networks, optimal learning in the case of feedforward networks, the local minimum is a real question in deeper neural learning and the case of Digital Dividing.**

Keywords: **Local Minima, Digital Divide, Exploitation, Exploration**

I. INTRODUCTION

A RNN is a class of artificial neural networks, which have a directed graph in a temporary sequence when connections between the nodes are created.

This helps it to show complex temporary behavior. Unlike neural feedback networks, the RNNs can process input sequences using their internal state (memory). The term "recurrent neural networks" refers to two broad categories of networks, with a similar general structure where one has a finite impulse and the other an endless impulse. This makes them applicable to tasks such as uncutting, linked handwriting(Li, Xiangang; Wu, Xihong2014).or speech recognition(Miljanovic, Milos. 2012), (Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J. 2009): Both network classes show dynamic temporal action. A recurring finite impulse network has an unrollable, direct acyclic characteristic which can be replaced by a neural network which is strictly feedbacked while a cyclical diagram which cannot be roll-out is a guided cyclical characteristic.

Recurring networks can have additional storage and storage via the neural network, with both final impulses and infinite pulsations. If the storage requires time delays and feedback loops, it can also be replaced with another network or graph. Such controlled states are called gated or gated memory, and below are long-term memory networks and intermittent gated systems. These controlled states are called gated states.

Regional minimum in certain neighbourhoods applies to a minimum and may not be a global minimum. Regional error function minima do not suit the lowest error value but only the lowest error in a limited set of independent variables ' values.

Gradient descent algorithms depend heavily on initialization to avoid minimum local levels. In particular for neural networks, it is a situation in which a neural learning network often occurs, in which weight changes for one or more patterns of the training simply counteract the adjustments to an already learned model.

Artificial intelligence has been in the background for decades, potting dust away, but the era is never over. In 2017, the dust cloud was disrupted and artificial intelligence entered a major trend. In a wide range of applications (for example, classification, approximation,(Frasconi, p., Gori, M., and Tesi A. 2000): signal processing, and pattern recognition (Fukumizu, K., Amari, S. 1999): Artificial Neural Networks (ANNs) are commonly used to improve neural networks ' output by learning from their environment. When learning takes place, the network's connection weights are adjusted without prior knowledge using a training data set.

Backpropagation (BP) algorithm is the most frequerid approach used to learn how to control neural feedforward networks (FNNSs). In relation to network modifiable weights, the BP algorithm measures the network error gradient. The BP algorithm will lead to movement to the minimum local level.

With regard to certain sequential machine learning functions, such as speech recognition, the predictive accuracy of RNNs is achieved time and again and no algorithm can equal them, although the first generation of RNNs were not that hot back in the day. They suffered from a significant setback from their error fixing cycle that kept their development up over the past decades. In the late 90s, a major breakthrough came into being that led to the creation of a new generation of RNNs that had been far more reliable for almost 20 years.

II. LOCAL MINIMA IN NEURAL NETWORKS

Controlled learning with a standard learning algorithm of a multi-layered neural network is facing the minimum local problems. In this condition, traditional training of the Neural Networks also involves changing the network weight of the link using a training collection of input output pairs without previous knowledge and adjusting the weights of the gradient downwards after a local error surface tendency leading to undesirable spots or a local minimum. Various (Toh, K.A. 2003): (Sprinkhuizen-Kuyper, L.G., Boers, E.J.W. 1999): (Miljanovic, Milos. 2012): addressed this issue by illustrating the nature of architecture and the local minima-free learning setting. In order to understand the behavior of the error surface, various types of local minima are studied in (Ferrari, S., Stengel, R.F. 2005): The local minimums are primarily related to two factors: the learning style and the network structure. A deterministic approach or probabilistic approach may be used for handling the problem. In an approach of determinism, instead of the primary gradient descent nodes, a new learning algorithm was introduced, global descent (Frasconi, p., Gori, M., and Tesi A. 2000):

Optimization algorithms are implemented to prevent local minima during the learning processes (Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J. 2009): but they require time to find the optimal worldwide. Another option, a probabilistic approach, is often a collection of weights such as the methods of initialization of weight (Jordanov, I.N., Rafik, T.A. 2004): which can minimize the likelihood of local minima being reached. Another fascinating approach is the neural network learning, where more than once training takes place, beginning with a random set of weights. The most suitable neural network is often chosen to describe the structure of the sun spot neural network model as the one with the least error, e.g. parquet (Ferrari, S., Stengel, R.F. 2005). Their best network was chosen from 10 architecturally similar networks, each with a random weight collection.

But it is not possible to specify the number of the random launch. The probabilistic approach to calculate the number of random starting times needed to complete neural network training was to avoid rebooting the time lyer and rhinchart(Goodfellow et al 2014) although it is easy to implement, but training takes a long time. An EANN can definitely be developed in yao, and it is demonstrated (Ferrari, S., Stengel, R.F. 2005): and(Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J. 2009): The comparison results are provided. This approach is successful in seeking a global minimum but still requires significant resources.

a. How Troublesome Are Spurious Local Minima in Deep Networks

For many years, deep learners blamed all their difficulties in training deep networks on bogus local minimum conditions. 1.1 Why Pesky's are Spurious Local Minimum at Deep Networks. Today, the question remains whether fake local minima with a high error rate compared to the global minimum are common in deep realistic networks. Nevertheless, a large number of recent studies suggest that most local minimums have error rates and general characteristics very similar to global minimum values.

One way we could naively address this problem is to map the value of the error function over time while we are building a deep neural network. Nonetheless, this technique does not provide sufficient information on the error surface, because it is hard to tell if the error surface is "humpy" or if we just have trouble determining the direction to which we should travel.

In 2014, Goodfellow et al. (a research team collaborating with Google and Stanford) created an article aimed at distinguishing these two possible confounding factors in order to examine the issue more effectively. They investigated cleverly what is happening on the error region between a randomly initialized parameter vector and a good ending choice by using linear interpolation instead of analysing the error functions over time. Given a randomly initialized vector parameter= I and stochastic gradient descent(SGD) solution + f, at every point along the linear interpolation= α • f+ (1-α) • • I We aim to compute the error function in every stage.

In other words, they wanted to investigate whether local minimum criteria could hinder our gradient search approach even if we knew in which direction. We showed that the direct path between a randomly initialized point in the parameter space and a stochastic gradient descent solution does not suffer from troubling local minima for a number of functional networks with different types of neurons.

In Feedforward networks, various researches have carried out several optimal learning analyzes to try and identify examples of local minimal conditions, which require a minimum of error on local surfaces.

III. OPTIMUM LEARNING IN FEED FORWARD NETWORKS

The form of the cost function depends on various parameters that make their effect difficult to identify. A quite interesting distinction is the dependency, in terms of network and learning environments, on the choice of the particular cost and feature, and on the structure of the problem. The former type of dependency will generate false local minima (Sak, Hasim; Senior, Andrew; Beaufays, Francoise 2014), (Frasconi, p., Gori, M., and Tesi A. 2000):while the latter type may lead to structural local minima (Sak, Hasim; Senior, Andrew; Beaufays, Francoise 2014), (Jordanov, I.N., Rafik, T.A. 2004): spurious local minima are independent of the issue to be solved and due to error input values above the cost function targets. This form of local minima is shown in (Fukumizu, K., Amari, S. 1999): for example. The problem of fake local minimum costs, for example the implementation of ordinary quadratic costs, quash functions and asymptotic goals (for instance,(Cantu-paz, E., Kamath, C. 2005),(Sexton, R.S., Gupta, J.N.D. 2000): or the use of LMS cost functionsminima (Sak, Hasim; Senior, Andrew; Beaufays, Francoise 2014), is sufficient to avoid these local minimum. This can easily lead to a mutual choice of costs or squashing functions to prevent problems of fake local minimas, as is the case.

In some cases it was investigated the structural reliance on network and data in order to identify optimal conditions that guarantee the absence of minimum levels. There can be a general analysis of the problem (Toh, K.A. 2003), which provides some theoretical conditions to ensure that pyramid network surfaces are locally free of minimal error.

There were also efforts to make the criteria to ensure the absence of territorial limits more clear. A number of inputs and many hidden unit networks are expected to achieve the backpropagation convergent. In (Toh, K.A. 2003): the research referred to above is particular for linearly separable patterns, which are likely to be present in many coordinate patterns, and therefore in many input networks. Networks with one hidden layer and as many hidden units as pattern (Toh, K.A. 2003): are expected to ignore local minima, however (Goodfellow et al 2014)

There is no limitation on the network architecture and the number of hidden units with the theory of a linearly separable sequence. One can argue that a one layer network is adequate to isolate the learning set in the case of linearly separable patterns. The relevance of the previous result however stems from a clear description, under that assumption, of the stationary points of the cost function. In fact, a learning algorithm is still able to find the optimum solution, unlike one-layer networks when the patterns are not linearly separable. As a consequence of this, we can see the linear separation as a limit over which problems for any learning algorithm are likely to begin. In the case of linearly-separating patterns, it should also be noted that the generalization of networks with a hidden layer to new examples is stronger than simple networks with a single layer. The weakness of the sources is the situation itself that certain pattern recognition issues, but certainly not in general, can be overcome

Recently, bianchini et al, have analyzed the problem of optimal learning in radial basis, function networks (Ferrari, S., Stengel, R.F. 2005). Under the assumption of patterns separable by hyperspheres, they have proven that attached cost function is local minima free moreover, they have also established some intriguing comparisons with hybrid learning, that is based on a self-organization first step for learning the weights of locally-tuned processing units and LMS for the output weights. A potential problem, which is likely to affect practical applications, is the learning process may be seriously plagued by the presence of local minima in the cost function. In general, there is a reason to exclude the presence of stationary points that may also be local minima obviously, this does not means that no learning procedure can effectively find optimal solutions, but, if the cost function has many local minima, devising an effective learning algorithm may become very difficult the presence of stationary points, and particularly of local minima, is something which affects more or less any algorithm.

IV. LOCAL MINIMA

Often known as relative minima are local minima. Regional minimum corresponds to a minimum in some places and may not be a total global minimum. Local error minimums do not match the lowest error value but only to the lowest error in a specific set of values of the independent variables. Local minimum error functions

In particular, in relation to neural networks, a learning neural network is sometimes implemented, where changes of weight to one or more training patterns simply compensate for the adjustments made to a previously-trained motif. Gradient downward algorithms strongly rely on their initialization to avoid local minima. It is not ideally suited for output mapping but is set in a less than optimal "locally" response mapping, which is called local minimum. The connection weights (weight jogging) can sometimes be prevented by a random jog. The basic definition of a locally-based minimum is often discussed in the plural: local minima. This means that the lower is the more correct or ideal solution.

Naturally, this leaves many loose ends as adaptive systems are often intertwined with philosophically unsolved problems as you start thinking about things like correctness.

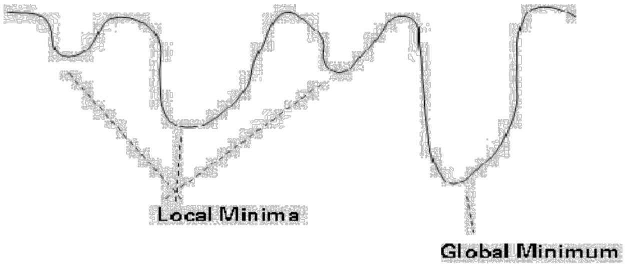


Fig. 1. Local minima and Global Minimum

a. *Is the Local Minima a real issue in deep neural learning?*

The primary challenge in optimizing deep learning models is that we are forced to use minimal local information to infer the global structure of the error surface. And, this can be a problem if weencounter a local minimum. Let's explain by taking an ant example who is trying to reach the global minima.

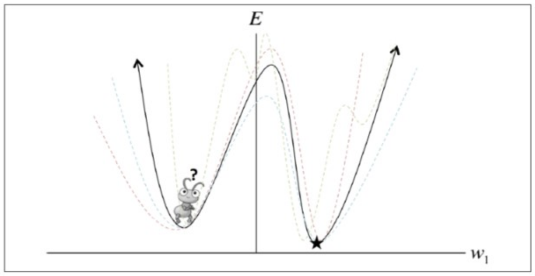


Fig.2. Ant Stuck to a deep local minima

As we see in the figure above, the ant attempts to reach the minimum low point (star here) to a point that it presumes to be the lowest point due to the lack of information concerning global data. Here, a local minimum point is the current status of ant.

Regional minima may potentially create a major problem, as it may lead to a poorly trained model. In practice, however, we can measure the effects of these local minimums on the model training process, just as the local minimums in the error area of deep education are normal.

The first question was answered by observation that the error surface is expected to be large and in some cases to have an infinite number of local minima. The first local minima source is related to the concept known as model recognition.

In sum up, in the profound learning model, the non-identifiable property results in many local minimas. Nonetheless, these local minima are necessarily not troublesome, although there are a large number of them.

This is because all unidentifiable configurations are compatible regardless of the input values fed. In the preparation, testing and evaluation dataset, all configurations would make the same mistake. In other words, from the training dataset all these models should learn the same thing. Actually, when they are fake, local minima are problematic.

The existence of the saddle points could also hinder the learning pace. The flat area of the errors surface is saddle dots, which almost zero makes the gradient value. It hinders the pace since there is ambiguity in the path to minimums. However, the presence of saddle points decreases exponentially with increasing size of the training data. Such flat areas of the error surface tend to be distracting. It should be noted here that any kind of optimisation in the second order would certainly hit these saddle spots.

The general form of local minima are the falsely and the error points that do occur in the model learning process present no potential threat.

One thing we should note is why so many advanced optimization algorithms were implemented (Adagrad, RMSProp, Adam). The explanation for this growth is that it seems to be the crucial challenge when examining the error surfaces of deep networks when optimizing deep networks is to find the right path forward. The gradient could change under our feet when we move, as a result of taking steps in the direction of global minimum levels. These advanced methods of optimization of the second order are therefore especially designed to tackle this problem.

V. DIGITAL DIVIDE

As we can see from the above example, the left hand image is kept to a degree that she wrongly believes is the best point due to a lack of digital information. The picture is a picture on the left side. Here, in the left picture, the current state of people is the local minimum level.

The ant, as can be seen in this figure, attempts to reach the lowest point (star in this case) to a point where, owing to lack of information on global data, it is ostensibly the lowest point. This is a local minimum level, the current state of ant.

Fig. 3. Digital divide

VI. LOCAL MINIMA, SADDLE POINTS, AND PLATEAUS

Different algorithms, such as those used for enhancement learning or to minimize functions often distinguish exploitation from exploration. The basic idea is that you have two kinds of actions you can do when you try to accomplish a task-good activities because you know they will produce good results and good ones, because you will get more knowledge about the task you try to accomplish.

Imagine sitting in one of your favorite restaurants to decide what to order. One choice is to buy your favorite dish-you still order the same thing. This is a good option because you're very confident you get something like it-we call it because you're using your previous menu details to produce a good result. Another way would be to order something new from your previously unknown menu. In the best case, something could end up better than your current preference, but something bad could also end up with you. While it might seem like a waste, if you end up in something worse, something new is still to be tried, because at least now you know it. You know your current option is more likely to be the best choice on the menu. You know you were never without something.

There is always a balance between the two approaches when making a number of decisions. Too much manipulation means that you can go through good opportunities in favor of sub-optical decisions you already know the outcome of, but being too exploratory means you may go through good chances in favor of other decisions that have less certainty about outcomes!

VII. EXPLOITATION

The minimization method (the job of defining the input into a math function generating a smallest output) is to hold the symptom that you are too exploitative in the so-called "local minimum," where you can not enhance the early non-optimic outcome by moving down a path to no return. On the other hand, the symptom of being excessively exploratory is a' slow convergence rate,' which takes a long time for the solution to be found, because it takes time to explore bad options with a little chance of better results than currently.

Philosophers, mathematicians, and economists also reflected this trade in human decision making in real life. We can see "exploratory" people who gradually come to work and seek to understand the subject in depth and thoroughly through all references and sub-references. These are individuals who are easy to distract and have trouble doing the final steps needed to get to the end of a project during the final stages. On the contrary, we see "exploitative" people as being far more goal oriented and often able to reach their goals rapidly but with a propensity to iterate towards the correct first solution, a behavior that often causes unintended drawbacks later in.

I think everyone can agree that adaptivity is crucial-that if you're persuaded it's going to be a good move to take action, while it's better to explore options that are less likely to produce good performance-but other than the best overall strategy to balance this trade-off? It's a well-explored area, mathematically speaking, with many academics working on thought experiments like the multi-arm bandit problem that pose a type of this puzzle, but there is still no general agreement.

And while there is clearly a always-needed balance between exploitation and discovery, a fascinating (in some sense profound) point is actually found which is often ignored in the more philosophical consideration of this topic.

According to this claim (explained below) a more exploitative policy is almost always a better choice than an exploratory policy in cases where there are numerous different possible actions, for example in many real world tasks. For understand why we first have to look at a recent theory of how saddle points in high dimensional spaces manifest themselves.

VIII. EXPLORATION

The name of the points on the "minimization landscape" which are local minimum points on one axis but not others is used to the mathematical problem of feature minimization saddle points. Intuitively, if we consider role minimisation as analogous to the concept of completing a task in the real world, saddle points can be thought of as points in which one form is no longer able to enhance the results of a task. Saddle point is not like local minima because they are not' dead ends,' and when you consider them' exploitable' methods may still be effective-as long as they have the flexibility to get off the axis that has reached the lowest level and to travel along the axis that still needs to be improved.

The theory itself notes that saddle points are significantly more prone to occurring in very large spaces (like the ones we would use to describe tasks with a large number of choices) than local minima. More precisely, only as you get closer to the true global minimum is the chance of local minima.

Another way of thinking about it a little more intuitively is something like this: if you are doing a job with a great many potentials, it is extremely unlikely that you will encounter a condition in which you really can not change, because if you are in this position, any action you can do will cause the situation to become worse than the one in which you currently live. Yet with the increase in the number of possible actions that you can take, given that every action is autonomous and special, the case alone becomes increasingly unlikely. When you truly find no change in any of your acts, this is probably an indication that you are close to the actual optical outcome, since this is the only case where such condition is possible.

So why does this hypothesis have an influence on the philosophical discussion of the equilibrium between explorative and exploitative behavior? Okay, it's because the main reason for more exploratory strategy is that local minima is prevented-a remedy that can trap you in a difficult situation that you can't escape later. Nevertheless, if local limits exist in the first place, they may lead to less successful exploratory behaviour.

In fact, because exploratory policies frequently do everything in their power to avoid local minima, people who use them could have developed a distorted view of the actual incidence rate of these minima–many of these local minima, which could be disguised as saddle points by exploratory policies. In other words, many of the solutions which the policy actively sought to avoid were hurried and botched together, in fact, subject to subtle, secret corrections or clauses that could have been obvious only when reached.

There is an additional danger that can have a significant effect on the excessively careful conduct of exploratory policies in high dimensional spaces when it comes to local minima (Policies that seek to move between greater exploratory actions and more exploitational behaving, depending on the cost landscape curvature). This risk is plateaus-sections of the landscape of optimization where the gradient is very low or slightly ondulating and noisy.

In the real world we can find plateaus as being analogous to situations in which the outcome tends to be very little change whatever action is taken, with an obvious noisy or random aspect to all observations.

The risk of plateaus is that the high, noisy gradient is likely to slow down adaptive policies and start exploring further to see whether they can find a more optimal way or at least collect more information on the effects of their actions. This can make sense, because plateaux can in many ways imitate local minima, and it is more exploratory to be trapped in local minima.

Nevertheless, perhaps in the opposite direction, speeding up and moving across the plateau as quickly as possible in the hard direction you assume it will be sloping until you are able to find a smoother surface that can provide you with details on the best behavior possible.

Again, this would seem to suggest that a more exploitative approach can work in many ways in high dimensional spaces better than expected.

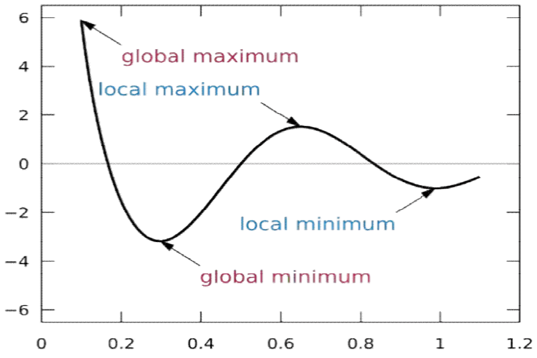


Fig. 4. Local and Global Minimum

As is the case in any attempt to apply wavy hand mathematical analysis to the real world, it is important to use a large grain of salt for this analysis and also to note that this applies only where a number of options or unknowns are involved. It makes no sense for example to apply it to our above menu example, because there are quite a few options-in this case an important approach such as refusing menu items that you obviously don't like and then attempting the rest thoroughly makes far more sense rather than simply picking and adhering to one item.

But there is certainly something to say because of your trust, focus and courage in taking big steps in your life, steps that might be dangerous or short-term. If anything, it's reassuring that we may not have to be as scared as we're all local minima and, despite the uncertainty enough, there's still a way out of any situation. When we think that this math theory applies to real life, it might also tell us to stay focused and move quickly toward our objectives. They are not trustworthy to follow our tricks until it has been seen. It seems sluggish and unstable to accelerate rather than slow. To always search for directions that can return better than the current one. We are not sure if we are careful to plan ahead for situations. However, note that it can actually be a good thing to be in a local minimum–it can mean that we are close to the real ideal solution.

The general theory of high dimension saddle points is a little like the old saying "as a door closes another door opens"–the concept is that while dead ends in existence can always be reached, chances are always present if new directions are to be followed.

The error surfaces are not convex in many advanced ML models, which implies that the descent gradient converges in sub-optimal valleys or to a local minima with all probability. Researches are being carried out to clarify the geometry and topology and to get better minimum levels out of these techniques. (Many of them have a high chance of "bumping" one out of poor local levels due to stochastic gradient descent).

IX. LOCAL MINIMA IN THE ERROR SURFACES OF DEEP NETWORKS

The main challenge in optimizing deep learning models is to use minimal local information in order to determine the global structure of the error surface. This is a difficult problem, because the local and the global systems generally do not correlate very much.

X. MODEL IDENTIFIABILITY

Local minima is the first source tied to a definition that is commonly known as model ID. One claim on deep neural networks is that their error surfaces have guaranteed an infinite number of local minima in some cases. This observation is valid for two main reasons.

The first is that any rearrangement of neurons within a completely connected feed-forward neural network will always give you the same final output at the bottom of the network-local minima only when it is fictitious. A falsely local minimum corresponds to a neural network weight configuration which makes a greater mistake than the global minimum configuration. If these types of local minima are normal, we quickly encounter major problems using gradient optimization techniques, since we can only take local structure into account.

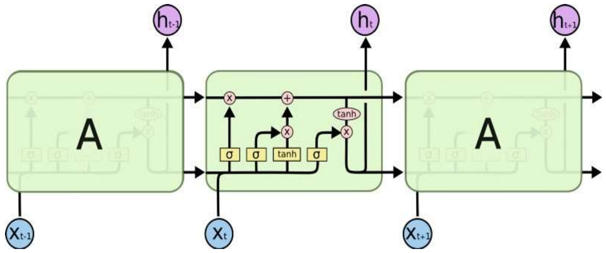
XI. WHAT IS THE PROBLEM IN NEURAL NETWORKS WITH LOCAL MINIMUM?

Several approaches have been suggested to solve the local minimum problems. A widely used one is more than once the training of a neural network starting with a random set of weights(Ferrari, S., Stengel, R.F. 2005). One advantage of the method is that other learning algorithms are easy to use and implement. Nevertheless, it takes longer for the networks to be educated. The neural networks function approximation is a non-convex optimization problem and there are therefore many local minima.

It is necessary, in most cases, to avoid the local minimum issue. If you have a low level of locality, however, then some of these methods could be tried:

11.1 The Big Breakthrough: Long Short Term Memory

Finally, the major breakthrough in the last 90s solved the downward flush problem and gave recurring network development a second wind. Long-term memory modules (LSTM) were at the core of this new approach.



As strange as the sound is, LSTM made a difference in the field AI. The time is long and short. Such new units or artificial neurons remember their inputs since a time ago, like the normal short-term memory units of the RNNs. Nevertheless, LSTMs will accumulate on their memories that have read / write properties similar to memory registers on a standard computer unlike regular RNN devices. Nevertheless, instead of digital, LSTMs have an analog, differentiable memory. In other words, its curves are continuous and its slopes are steep. We are therefore well adapted to the partial differential calculus involved in the rear and gradient decrease.

LSTMs are not only able to tweak the weight of the data they have stored. In addition, they can hold, erase, transform and otherwise monitor the inflow and outflow through their training. LSTMs will mainly add substantial misinformation for sufficient time to sustain relatively steep gradients and thus a relatively short training duration. It removes the issue of the absence gradient and greatly enhances the exactness of the recurrent LSTM networks today. Due to this incredible progress in RNN's architecture, the core of their businesses now uses RNNs for power applications, as well as many other leading companies, not to mention start-ups. All of a sudden, RNNs are a big deal.

11.2 Gradient Descent

One of the most popular gradient descent algorithms in machine learning is known. His greatest virtue is his remarkable ability to sidestep the dreaded' dimensionality curse.' This problem plagues systems with far too many variables, such as neural networks, to permit the estimation of the brute forces with optimal levels. Gradient descent removes however the dimension curse by closing the multi-dimensional error or cost function to the local lowpoint, or local minimum. It allows the system to assess the change value or weight, which will assign accuracy back to each of the network's units.

11.3Enhanced learning rate: if your algorithm's learning rate is too low then it will be held to a local minimum.

11.4 Expanding hidden layers / units: it could increase the approximation of the feature.

11.5 Try various activation functions: Make sure the model and data set are appropriate for the combination of activation functions.

11.6 Trying various optimization algorithms: Consider using algorithms such as Adam's Optimizer and RMSProp instead of a traditional gradient descent.

XII. CONCLUSION

Neural networks or RNNs may recall their earliest inputs, which gives them a wide margin in sequence, context-sensitive tasks such as spoken recognition over other artificial neural networks. These do not refer to the lowest error value local minima of the error function but only to the lowest error within a limited range of independent variables values. Gradient descent algorithms rely strongly on their initialization so that local minima are avoided. In neural networks, it is a phenomenon that sometimes a learned neural network gets into it, where the weight changes for one or more training patterns merely compensate for the modifications made to a previously trained model. It is not the optimal output mapping of the previously trained pattern but is trapped in a lower than ideal "local" response mapping called local minimum.

12.1 Recommendation

In order to effectively use the new design, the recommendation is as follows: è the hardware and software requirement should be specified.

i. To work effectively and achieve the maximum results, the Workers should have basic computer operating knowledge.

ii. To avoid local minima, more research and theory should be developed.

12.2 Suggestions for further studies

The investigator suggests that all researchers who wish to do further work on how to strengthen or reduce the problem with the local minima system should do more research on the use of discrepancies in reporting systems between local minimum and global minima.

REFERENCE

[1] Cantu-paz, E., Kamath, C. (2005): *"An Emperical Comparison of Combination of Evolutionary Algorithm and Neural Networks for Classification Problems".* In: IEEE Transaction on Systems, Man and Cybernetics, Part B, 915-927

[2] Ferrari, S., Stengel, R.F. (2005): *"Smooth Function Approximation Using Neural Networks".* In: IEEE Transactions on Neural Networks, 24-38.

[3] Frasconi, p., Gori, M., and Tesi A. (2000): *"Successes and Failures of Backpropagation: A Theoritical Investigation,".*Chapter in Progress in Neural Networks, Ablex Publishing, Omid Omidvar (Ed.), in press

[4] Fukumizu, K., Amari, S. (1999): *" Local Minima and Plateaus in Multilayer Neural Networks".* In: 9th International Conference on Artificial Neural Networks, 597-602

[5] Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J. (2009): *"A Novel Connectionist System for Improved Unconstrained Handwriting Recognition"(PDF*). IEEE Transaction on Pattern Analysis and Machine Intelligence.31(5):855-868. citeseerX10.1.1.139.4502. doi:10.1109/tpami. PMID 19299860.

[6] Jordanov, I.N., Rafik, T.A. (2004): "*Local Minima Free* Neural Networks Learning". In 2nd International IEEE Conference on Intelligent Systems, 34-39

[7] Li, Xiangang; Wu, Xihong (2014). *"Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition*". arXiv:1410.4281 [cs.CL].

[8] Miljanovic, Milos. (2012): "*Comparative Analysis of Recurrent and Finite Impulse Response Neural Networks in Time Series Prediction".* (PDF) Indian Journal of Computer and Engineering. 3(1).

[9] Sak, Hasim; Senior, Andrew; Beaufays, Francoise (2014). *"Long Short-Term Memory recurrent neural network architectures for large scale acoustic modeling" (PDF).*

[10] Sexton, R.S., Gupta, J.N.D. (2000): "*Comparative Evaluation of Genetic Algorithm and backpropagation for Training Neural Networks".* In: Information Sciences, 45-59

[11] Sprinkhuizen-Kuyper, L.G., Boers, E.J.W. (1999): *"A Local Minimum fo the 2-3-1 XOR Network"*. In: IEEE Transactions on Neural Networks, 968-971

[12] Toh, K.A. (2003): "*Deterministic Global Optimization for FNN Traning*". In: IEEE Transactions on Systems, Man and Cybernetics, part b, 977-983

[13] https://universalflowuniversity.com/Books/Computer%20Programming/Machine%20Learning%20and%20Deep%20Learning/Fundamentals%20of%20Deep%20Learning_%20Designing%20NextGeneration%20Machine%20Intelligence%20Algorithms.pdf

[14] https://stats.stackexchange.com/questions/203288/understanding-almost-all-local- minimum-have-very-

[15] Goodfellow et al (2014)"*Avoididing local minima problem in backpropagationalgorithem"* In: IEEE Transactions on Systems on neural networks.