Second Order Optimization Methods in Feed Forward Neural Nets with a Knowledge Based Expert System Control Loop

Simon Michael Herrmann[1,2], **Hans Ehm**[2]
[1]Technical University Munich, Department of Mathematics, Munich, Germany
simon.herrmann@tum.de
[2]Infineon Technologies AG, Am Campeon 1-15, 85579 Neubiberg, Germany
simonmichael.herrmann@infineon.com & hans.ehm@infineon.com

Abstract— *The scope of this extended abstract is to introduce two connected ideas and encourage a discussion. Firstly, we present second order methods equipped with a new approximation on the Hessian matrix in the training phase of a feed forward neural net. Secondly, we pitch the idea to combine knowledge based expert systems and neural nets to establish confidence in the predicting power of neural nets.*

Keywords: Second order optimization, hybrid system, artificial feed forward neural network, knowledge based expert system

1. Introduction

As a way of statistical modelling, feed forward neural nets (FFN) have become an important field of research as access to vast amounts of data rises and computational resources have gained the power to handle them. We define a FFN as a function $f : \mathcal{X} \subset \mathbb{R}^m \times \Theta \to \mathcal{Y} \subset \mathbb{R}^n$ where Θ constitutes the parameter space, $m \in \mathbb{N}$ the input dimension, and $n \in \mathbb{N}$ the output dimension.

Firstly, we focus on the training phase, particularly on the empirical risk minimization (ERM) problem introduced in [1]. The objective function $R(\theta)$ of the ERM is a composition of highly nonlinear loss functions. The dimension of the parameter space Θ grows with the number of neurons in a FFN, a deep FFN is linked to a high dimensional minimization problem. We observed that the objective function $R(\theta)$ covers large areas with small or no slope, a phenomenon that occurs with commonly used activation functions. A great challenge is finding an exact global solution of the ERM problem. Thus, we focus on solving the ERM problem approximately with the help of sophisticated non linear optimization techniques. Iterative algorithms are a favoured approach to find approximated local solutions, denoted by $\hat{\theta} \in \Theta$.

The second topic is aimed at questioning the user's confidence in FFNs. Practical experience revealed low user acceptance of computer-aided decision-making when the underlying decision support system is perceived as a black box. We presume, as displayed in [2], that the user may not have knowledge of the underlying decision support system, nor the details of how the system arrives at its prediction. We want to extend the user's confidence in its willingness to act on the basis of the recommendations, actions, and decisions of an artificial intelligent decision support system [2].

2. Background

2.1 ERM Problem and Second Order Methods

Second order optimization describes a class of iterative algorithms which use curvature information of the underlying problem to create a decreasing sequence in order to solve a minimization problem. This additional information, compared to first order methods, capacitates a faster convergence of the decreasing sequence to a local minimum using greater step sizes. Therefore, descent sequences created by second order methods are able to overcome areas of small slope faster than sequences created by first order methods [3]. The curvature information of the objective function is contained in the Hessian matrix of the objective function. As second derivative of the $R(\theta)$ with respect to θ, the size of the Hessian matrix increases with the number of neurons. Therefore, highly accurate, easily calculated and efficiently stored approximations of the Hessian matrix are required. The Kronecker factorized approximation of the curvature matrix (KFAC) constitutes a novel approach [4]. The KFAC, an approximation of the Hessian matrix using probabilistic structures within the problem, requires less memory and is calculated faster. The KFAC-algorithm in [4] is an approximated version of the natural gradient method by Amari [5]. A second order optimization algorithm, called the Hessian-Free (HF), was introduced by Martens et al. [6]. The HF is a Newton-like method which contains a precondition conjugated gradient method (PCG) as sub routine. In test runs, these algorithms outperformed first order methods, e.g. the stochastic gradient method, in both running time and accuracy [7].

2.2 Comprehensibility

Only limited data is available regarding our investigation, i.e. establishing confidence in the predictions of FFN, as the way a FFN derives its predictions is often difficult to comprehend. However, comprehensible predictions

are important, if the recipient is a human agent. In contrast, knowledge-based expert systems (KBES), defined as $g : \mathcal{M} \subset \mathbb{R}^m \rightarrow \mathcal{N} \subset \mathbb{R}^n$, are widely accepted among users, especially in high knowledge domains, including industrial applications like the complex but highly structured semiconductor environment [2]. We pitch the hypothesis that the acceptance of a neural net's prediction increases when a KBES is endorsing them.

3. Proposed Approach

The novel ideas are solving ERM problems faster and with higher accuracy utilizing a second order method as well as, using the, KBES to endorse the FFN's predictions.
In the field of second order optimization we started to adjust existing algorithms with the KFAC approximation [8].
The idea for the comprehensibility is to check if the FFN prediction is in- or outside the range of the KBES. Therefore, let the range of the KBES be a subset embedded in the range of the FFN e.g. $g(\mathcal{M}) \subset f(\mathcal{X}, \hat{\theta})$. For every output of the FFN $f(x, \hat{\theta})$ within the range of the KBES $g(\mathcal{M})$, we can find a set $D \subset \mathcal{M}$ such that $g(D) = f(x, \hat{\theta})$ (holds for the case $f(\mathcal{X}, \hat{\theta}) \subset g(\mathcal{M})$, too). The significance of the FFN's output $f(x, \hat{\theta})$ is enhanced by the KBES, e.g. $g(D)$. Given that the FFN's output is not contained in the subset $f(x, \hat{\theta}) \notin g(\mathcal{M})$, the supporting capability expires. In this particular case, an alert is triggered requesting a (human) agent interaction, e.g. to either accepting extended border parameters and change the range of the KBES or decline.

4. Method and First Results

As previously mentioned, the potential of the HF algorithm to solve ERM is confirmed by several examples [7]. After investigating the HF algorithm in detail, it appears that the PCG iterations are costly in terms of time. Keeping that in mind, we equipped the PCG sub routine in the HF with the KFAC approximation. Using the KFAC approximation as preconditioner in the PCG, we obtained a Hessian Matrix approximation which can be stored and calculated efficiently. The curvature information in the KFAC approximation allows greater step sizes in the sub routine and uses less iterations as a result [8].
With respect to the comprehensibility, research on hybrid systems with a KBES providing an expectance corridor for the output of a FFN has been conducted. At this point hybrid systems, where the KBES adjusts the FFN or vice versa, make up the majority of available research data [9].

5. Limitations and Outlook

The convergence theory of the KFAC to the Hessian matrix is not completed. Under special assumptions, the KFAC approximation of the Hessian matrix converges to the Gauss Newton matrix [8]. The Gauss Newton matrix is an often used approximation of the Hessian matrix [3]. While practical experiments have demonstrated the ability and potential of the KFAC, the limitation of examples should be kept in mind.
The challenge of the KBES is to find the correct parameters to build confidence in the outcome of the FFN, especially if $\dim(\mathcal{M})$ is strikingly smaller than $\dim(\mathcal{X})$. Another challenge is to identify the subset D, as described earlier, such that for all $y \in D$ the distance between x and y is minimal with respect to a metric.

6. Conclusion

In test runs, we were able to show that second order methods, equipped with the KFAC approximation, are capable of solving the ERM problem faster and with higher accuracy. The use of second order methods, especially with the help of KFAC, is promising and deserves further research. When investigating the training of a FFN linked to the ERM problem more closely, further efficient approximations might be found, similar to the KFAC as described e.g. in [10].
Fast and accurate predictions of a FFN itself are not sufficient enough to increase trust in applications providing output to human agents. The intention of this abstract is to enhance the amount of FFNs in high knowledge domain areas, where KBES can be used to support the confidence in the predictions of the FFN.

References

[1] V. Vapnik, "Principles of risk minimization for learning theory," in *Proceedings of the 4th International Conference on Neural Information Processing Systems*, ser. NIPS'91. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1991, pp. 831–838. [Online]. Available: http://dl.acm.org/citation.cfm?id=2986916.2987018

[2] M. Madsen and S. Gregor, "Measuring human-computer trust," in *Proceedings of the 11 th Australasian Conference on Information Systems*, 2000, pp. 6–8.

[3] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer, 2006.

[4] J. Martens and R. B. Grosse, "Optimizing neural networks with kronecker-factored approximate curvature," *CoRR*, vol. abs/1503.05671, 2015. [Online]. Available: http://arxiv.org/abs/1503.05671

[5] S. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, Feb. 1998. [Online]. Available: http://dx.doi.org/10.1162/089976698300017746

[6] J. Martens and I. Sutskever, *Training Deep and Recurrent Networks with Hessian-Free Optimization*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 479–535. [Online]. Available: https://doi.org/10.1007/978-3-642-35289-8_27

[7] R. Kiros, "Training neural networks with stochastic hessian-free optimization," *CoRR*, vol. abs/1301.3641, 2013. [Online]. Available: http://arxiv.org/abs/1301.3641

[8] S. M. Herrmann, "Training neuronaler netze mit newton-artigen verfahren," Technical University of Munich, Department of Mathematics, Bachelor Thesis, Advisor: Sebastian Garreis, Supervisor: Prof. Dr. Michael Ulbrich, 2017.

[9] A. Castro and V. Miranda, "Mapping neural networks into rule sets and making their hidden knowledge explicit application to spatial load forecasting," *14th PSCC*, 2002.

[10] A. Botev, H. Ritter, and D. Barber, "Practical Gauss-Newton Optimisation for Deep Learning," *ArXiv e-prints*, June 2017.