Palagi, Laura; **Seccia, Ruggiero**

Block layer decomposition schemes for training deep neural networks. (English)

Zbl 1441.90127
J. Glob. Optim. 77, No. 1, 97-124 (2020).

Summary: Deep feedforward neural networks' (DFNNs) weight estimation relies on the solution of a very large nonconvex optimization problem that may have many local (no global) minimizers, saddle points and large plateaus. Furthermore, the time needed to find good solutions of the training problem heavily depends on both the number of samples and the number of weights (variables). In this work, we show how block coordinate descent (BCD) methods can be fruitful applied to DFNN weight optimization problem and embedded in online frameworks possibly avoiding bad stationary points. We first describe a batch BCD method able to effectively tackle difficulties due to the network's depth; then we further extend the algorithm proposing an online BCD scheme able to scale with respect to both the number of variables and the number of samples. We perform extensive numerical results on standard datasets using various deep networks. We show that the application of BCD methods to the training problem of DFNNs improves over standard batch/online algorithms in the training phase guaranteeing good generalization performance as well.

MSC:

90C26 Nonconvex programming, global optimization
90C06 Large-scale problems in mathematical programming
68T05 Learning and adaptive systems in artificial intelligence

Keywords:

deep feedforward neural networks; block coordinate decomposition; online optimization; large scale optimization

Software:

AdaGrad; Adam; Keras; RMSprop; SciPy

Full Text: DOI

References:

[1] Beck, A.; Tetruashvili, L., On the convergence of block coordinate descent type methods, SIAM J. Optim., 23, 4, 2037-2060 (2013) · Zbl 1297.90113

[2] Bertsekas, DP, Incremental least squares methods and the extended Kalman filter, SIAM J. Optim., 6, 3, 807-822 (1996) · Zbl 0945.93026

[3] Bertsekas, DP, Nonlinear programming, J. Oper. Res. Soc., 48, 3, 334-334 (1997)

[4] Bertsekas, D.P.: Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. CoRR, arXiv:abs/1507.01030 (2015)

[5] Bertsekas, DP; Tsitsiklis, JN, Gradient convergence in gradient methods with errors, SIAM J. Optim., 10, 3, 627-642 (2000) · Zbl 1049.90130

[6] Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: COMPSTAT (2010)

[7] Bottou, L.; Curtis, FE; Nocedal, J., Optimization methods for large-scale machine learning, SIAM Rev., 60, 2, 223-311 (2018) · Zbl 1397.65085

[8] Bravi, L.; Sciandrone, M., An incremental decomposition method for unconstrained optimization, Appl. Math. Comput., 235, 80-86 (2014) · Zbl 1334.90071

[9] Buzzi, C.; Grippo, L.; Sciandrone, M., Convergent decomposition techniques for training RBF neural networks, Neural Comput., 13, 8, 1891-1920 (2001) · Zbl 0986.68109

[10] Chauhan, V.K., Dahiya, K., Sharma, A.: Mini-batch block-coordinate based stochastic average adjusted gradient methods to solve big data problems. In: Proceedings of the Ninth Asian Conference on Machine Learning, volume 77 of Proceedings of Machine Learning Research, pp. 49-64. PMLR, 15-17 Nov 2017

[11] Chollet, F., et al.: Keras (2015)

[12] Dauphin, YN; Pascanu, R.; Gulcehre, C.; Cho, K.; Ganguli, S.; Bengio, Y., Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, Adv. Neural Inf. Process. Syst., 27, 2933-2941 (2014)

[13] Defazio, A.; Bach, F.; Lacoste-Julien, S., SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives, Adv. Neural Inf. Process. Syst., 27, 1646-1654 (2014)

[14] Duchi, J.; Hazan, E.; Singer, Y., Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res., 12, 2121-2159 (2011) · Zbl 1280.68164

[15] Fisher, RA; Johnson, NL; Kotz, S., Statistical methods for research workers, Breakthroughs in Statistics, 66-70 (1992), Berlin: Springer, Berlin

[16] Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249-256 (2010)

[17] Goodfellow, I.; Bengio, Y.; Courville, A., Deep Learning (2016), Cambridge: MIT Press, Cambridge · Zbl 1373.68009

[18] Grippo, L.; Manno, A.; Sciandrone, M., Decomposition techniques for multilayer perceptron training, IEEE Trans. Neural Netw. Learn. Syst., 27, 11, 2146-2159 (2016)

[19] Grippo, L.; Sciandrone, M., Globally convergent block-coordinate techniques for unconstrained optimization, Optim. Methods Softw., 10, 4, 587-637 (1999) · Zbl 0940.65070

[20] Huang, G.; Zhu, Q.; Siew, C., Extreme learning machine: theory and applications, Neurocomputing, 70, 489-501 (2006)

[21] Huang, G-B; Wang, DH; Lan, Y., Extreme learning machines: a survey, Int. J. Mach. Learn. Cybern., 2, 2, 107-122 (2011)

[22] Johnson, R.; Zhang, T., Accelerating stochastic gradient descent using predictive variance reduction, Adv. Neural Inf. Process. Syst., 26, 315-323 (2013)

[23] Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: open source scientific tools for Python. [Online; Accessed \(<\) today \(>\] (2001)\)

[24] Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR, arXiv:abs/1412.6980 (2014)

[25] Nesterov, Y., Efficiency of coordinate descent methods on huge-scale optimization problems, SIAM J. Optim., 22, 2, 341-362 (2012) · Zbl 1257.90073

[26] Nesterov, YE, A method for solving the convex programming problem with convergence rate o \((1/{\rm k}\hat{}2)\), Dokl. Akad. Nauk SSSR, 269, 543-547 (1983)

[27] Nocedal, J.; Wright, SJ, Numerical Optimization (2006), New York: Springer, New York

[28] Palagi, L., Global optimization issues in deep network regression: an overview, J. Glob. Optim., 73, 239-277 (2018) · Zbl 1421.90154

[29] Qin, T.; Scheinberg, K.; Goldfarb, D., Efficient block-coordinate descent algorithms for the group lasso, Math. Program. Comput., 5, 6, 143-169 (2013) · Zbl 1275.90059

[30] Robbins, H.; Monro, S., A stochastic approximation method, Ann. Math. Stat., 22, 400-407 (1951) · Zbl 0054.05901

[31] Tieleman, T.; Hinton, G., Lecture 6.5-RMSProp: divide the gradient by a running average of its recent magnitude, COURSERA Neural Netw. Mach. Learn., 4, 2, 26-31 (2012)

[32] Wang, H., Banerjee, A.: Randomized block coordinate descent for online and stochastic optimization. arXiv preprint arXiv:1407.0107 (2014)

[33] Wright, SJ, Coordinate descent algorithms, Math. Program., 151, 1, 3-34 (2015) · Zbl 1317.49038

[34] Yu, A.W., Huang, L., Lin, Q., Salakhutdinov, R., Carbonell, J.: Normalized gradient with adaptive stepsize method for deep neural network training. CoRR arXiv:abs/1707.04822 (2017)

[35] Zhao, T., Yu, M., Wang, Y., Arora, R., Liu, H.: Accelerated mini-batch randomized block coordinate descent method. In: Advances in Neural Information Processing Systems, pp. 3329-3337 (2014)