Phd thesis proposal

Metaheuristics for deep learning architectures. Application on computer vision problems

Supervised by Prof. El-Ghazali Talbi (CRISTAL Laboratory CNRS, University of Lille and INRIA Lille Nord Europe)
Co-supervised by Dr. Nadiya Shvai (Vinci)

This last decade, deep learning methods have attracted an important interest for different applications as image recognition [1], speech recognition [2], and natural language processing [3]. The most successful methods have multi-level architectures where there is an alternation between layers of linear transformations and *max* function. For instance, the *max* functions are known as ReLUs (Rectified Linear Units) and compute the mapping $y = max(x, 0)$ in a pointwise fashion [5], in convolutional networks [3] and *maxout* networks [6], the *max* operation is performed over a small set of variable within a layer. The use of a supervised learning becomes nowadays a standard to train very deep networks. The objective function in most cases is the *cross-entropy*, that must be minimized using a stochastic gradient descent (SGD), in which the gradient is evaluated using the back-propagation procedure.

The general shape of the loss function is very poorly understood. In the eighties, most of researchers were turned to deal with relatively small networks, because the convergence tends to be unreliable when using batch optimization. Then, multilayer neural nets had a reputation of being finicky and unreliable, then, the community to focus on simpler method with convex objective functions (loss functions), such as kernel machines and boosting. However, some work experimenting larger networks and SGD showed that, while multilayer nets do have many local minima, the result of multiple experiments consistently give very similar performance. Then, the increase of local minima, facilitates their find, and surprisingly, they are all more or less equivalent in terms of performance on the test set.

Moreover, it was established in a recent work that the loss function of a typical multilayer net with ReLUs can be expressed as a polynomial function of the weights in the network, whose degree is the number of layers, and whose number of monomials is the number of paths from inputs to output. As the weights (or the inputs) vary, some of the monomials are switched off and others become activated, leading to a piecewise, continuous polynomial whose monomials are switched in and out at the boundaries between pieces.

Then, an important interrogation concerns the distribution of critical points (maxima, minima, and saddle points) of such functions. Loss surfaces for large neural nets have many local minima that are essentially equivalent from the point of view of the test error, and these minima tend to be highly degenerate, with many eigenvalues of the Hessian near zero.

The goal of the present thesis is to attempt to explain this property. It is also an attempt to shed light on the puzzling behavior of modern deep learning systems when it comes to optimization and generalization. Moreover, the advantage of metaheuristics do not compute gradients and thus do not tend to become trapped in high-index saddle points. Afterwards, as in case deep nets, the goal is not to find the global optimum on the training set because it may lead to overfitting. Consequently, metaheuristics seems to be suited to tackle this problem. These methods were not yet explored to optimize big problems (deep nets can have several millions of parameters) and the goal of this thesis is to design metaheuristics to solve this problem.

References

[1] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In NIPS

[2] Hinton, G., Deng, L., Yu, D., Dahl, G., rahman Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. Signal Processing Magazine.

[3] Weston, J., Chopra, S., and Adams, K. (2014). #tagspace: Semantic embeddings from hashtags. In EMNLP.

[4] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998a). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86:2278–2324.

[5] Nair, V. and Hinton, G. (2010). Rectified linear units improve restricted boltzmann machines. In ICML.

[6] Goodfellow, I. J., WardeFarley, D., Mirza, M., Courville, A., and Bengio, Y. (2013). Maxout networks. In ICML.

Company Vinci Autoroute is an important world wide company in traffic prediction and pattern recognition in the context of transportation systems (https://www.vinci-autoroutes.com).

Places of exercice of the thesis Paris and Lille (alternation to study)

Candidate's scientific profil Master or equivalent in computer science and/or applied mathematis perfectly mastering machine learning, optimization, and operational research.

Contacts : Send extended CV + master's degree + master notes semesters 1 & 2 + cover letter to: el-ghazali.talbi@univ-lille.fr