Substituting ReLUs with Hermite Polynomials gives faster convergence for SSL

Vishnu Lokhande Sathya N. Ravi Songwong Tasneeyapant Abhay Venkatesh
Vikas Singh
University of Wisconsin-Madison

This poster abstract is submitted for REGULAR poster presentation and STUDENT COMPETITION.

Abstract: Rectified Linear Units (ReLUs) are among the most widely used activation function in a broad variety of tasks in vision. Recent theoretical results suggest that despite their excellent practical performance, in various cases, a substitution with basis expansions (e.g., polynomials) can yield significant benefits from both the optimization and generalization perspective. Unfortunately, the existing results remain limited to networks with a couple of layers, and the practical viability of these results is not yet known. Motivated by some of these results, we explore the use of Hermite polynomial expansions as a substitute for ReLUs in deep networks. While our experiments with supervised learning do not provide a clear verdict, we find that this strategy offers considerable benefits in semi-supervised learning (SSL) settings. We carefully develop this idea and show how the use of Hermite polynomials based activations can yield improvements in pseudo-label accuracies and sizable financial savings (due to concurrent runtime benefits). Further, we show via theoretical analysis, that the networks (with Hermite activations) offer robustness to noise and other attractive mathematical properties.

From theory to practice: Recent literature that have been providing guidance on learning theoretic or optimization-centered properties of deep networks have only been limited to networks with a small (or in some cases, one or two) number of layers. For example, critical point and convergence analysis for two layered ReLU networks has been derived [1]. Next, the results in [2] give interesting insights into how over-parameterization using multi-layer feedforward ReLU based networks can help generalization for two layer networks. Within the last year, [3] investigated optimizing the population risk of the loss using stochastic gradient descent and showed for one hidden layer network, one could avoid spurious local minima by utilizing an orthogonal basis expansion for ReLUs. Relatively less is known whether these strategies are a good idea for the architectures in broad use in computer vision today. Our work tries to bridge this gap in literature by building up from [3], using the orthogonal basis expansion of ReLU's (Hermite polynomials) to mold the optimization landscape of deep networks.

SSL framework: One way to approach the SSL problem would be to run a two phase procedure. In Phase I, we may estimate pseudolabels for the unlabeled images. Then, in Phase II, using the estimated pseudolabels, we may train any standard classifier such as ResNets on the entire dataset, in a completely supervised manner [4]. The success of the two step procedure depends on both the quality of the pseudolabels and the ease of training a model (say, a ResNet) in the second step. We make use of networks with hermite activations to generate better quality of pseudolabels thereby improving the SSL training procedure. One of the reasons for the gains is a smoother optimization landscape, when using Hermites, since all the neurons are always active during training with probability 1.

Conclusions: We studied the viability and potential benefits of using a finite Hermite polynomial bases as activation functions, as a substitute for Rectified Linear Units (ReLUs). The lower order Hermite polynomials are known to have nice mathematical properties from the optimization landscape point of view, although little is known in terms of their practical applicability to networks with more than a few layers or to interesting tasks in vision. We observed from our extensive set of experiments that incorporating Hermite polynomials into the deep network architectures can yield significant computational benefits, and we demonstrate the utility of this idea in a computationally intensive semi-supervised learning task. Under the assumption that the training is being performed on the cloud and with published pricing structure, we show sizable financial savings are possible. On the mathematical side, we also showed that Hermite based networks have nice noise stability properties that appears to be an interesting topic to investigate, from the robustness or adversarial angles.

References

[1] Y. Tian, "An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3404–3413, JMLR. org, 2017.

[2] C. Wei, J. D. Lee, Q. Liu, and T. Ma, "On the margin theory of feedforward neural networks," *arXiv preprint arXiv:1810.05369*, 2018.

[3] R. Ge, J. D. Lee, and T. Ma, "Learning one-hidden-layer neural networks with landscape design," *arXiv preprint arXiv:1711.00501*, 2017.

[4] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Representation Learning, ICML*, vol. 3, p. 2, 2013.