Off-line Learning Analysis for Soft Committee Machines with GELU Activation

Frederieke Richert[1], Michiel Straat[2], and Michael Biehl[1]

[1]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands
[2]Center for Cognitive Interaction Technology, Bielefeld University, Germany

Abstract

Statistical physics has played an important role in the last decades in increasing the theoretical knowledge about the typical behavior of neural networks. One phenomenon that has surfaced recently is different phase transitions in off-line learning for two-layer feed-forward networks depending on the activation function used in the network. In a recent paper [1], it was shown that while the sigmoidally shaped erf results in a first-order phase transition when varying the training set size, the now commonly used ReLU activation function displays a second-order specialisation of hidden units in shallow networks. This can be of practical relevance, because in the case of a first-order phase transition a suboptimal solution remains stable after the transition point and training might get stuck in this solution instead of converging to the optimal solution.

In the search for activation functions which give an insight into the features of the functions that lead to the respective type of specialisation in the network, we investigated the well-known GELU activation function [2], $\mathrm{GELU}(\gamma, x) := x(1 + \mathrm{erf}[\gamma x/\sqrt{2}])/2$. It is a smooth approximation of the ReLU, as $\lim_{\gamma\to\infty} \mathrm{GELU}(\gamma, x) = \mathrm{ReLU}(x)$. Furthermore, GELU approximates the currently often used swish [3].

In the statistical mechanics approach to off-line learning the assumption is that the training process optimises the network with respect to the whole entire set of example data. Eventually the configuration of the network in the parameter space can be described by a Gibbs-Boltzmann distribution under certain simplifying assumptions. Learning curves can be obtained in this framework by optimising the free energy of the network for a given relative training set size α. The generalisation error drops to a plateau value for small values of α. Above a critical α-value, the generalisation error either jumps to a lower value or falls off with a changed slope. These are signs of first or second order phase transition respectively.

The type of phase transition becomes especially evident by observing the so-called order parameters, which represent the overlaps between weight vectors of different hidden units $Q_{ik} = \mathbf{w}_i \cdot \mathbf{w}_k$ and $R_{in} = \mathbf{w}_i \cdot \mathbf{w}_n^*$. Here, $\mathbf{w}_i$ is the weight vector of the i-th hidden unit of the student network and $\mathbf{w}_n^*$ is the n-th hidden unit of the teacher network. The phase transitions lead to specialisation among the order parameters, so that some increase

rapidly after the transition point and others decrease. Again, a discontinuity in the order parameters corresponds to a first-order phase transition and a continuous evolution to a second-order phase transition.
From the statistical mechanical analysis of a soft committee machine with GELU activation function, involving the derivation of the generalisation error and the optimisation of the corresponding free energy, we find that the GELU network shows a second-order phase transition independent of the slope parameter γ and the hidden layer size K. This is obtained by deriving learning curves for different values of γ and K. The results for different K motivate to investigate also the case of an infinitely large hidden layer, $K \to \infty$. In this analysis, it is easy to see the independence of the type of the transition from γ, but that there is indeed an optimal γ for which the α at which the transition occurs is minimal.

These results for the GELU can be related to the previously obtained results for the erf and the ReLU and support the assumption that the nature of the phase transition of the GELU should be the same as for the ReLU. The hypothesis that non-convex activation functions result in first-order transitions is ruled out, due to the non-convexity of the GELU. Furthermore, the results illuminate why not only the ReLU but also the swish are currently such popular choices for activation functions.

References

[1] E. Oostwal, M. Straat, and M. Biehl. "Hidden unit specialization in layered neural networks: ReLU vs. sigmoidal activation". In: *Physica A*, Vol. 564, 125517 (2021). DOI: 10.1016/j.physa.2020.125517.

[2] D. Hendrycks, and K. Gimpel. "Gaussian Error Linear Units (GELUs)" In: *arXiv e-prints*(2016). DOI:10.48550/arXiv.1606.08415

[3] P. Ramachandran, B. Zoph, and Q.V. Le. "Searching for activation functions." In: *6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, workshop track proceedings*