Learning on tree architectures outperforms a convolutional feedforward network

Yuval Meir¹, Itamar Ben-Noam¹, Yarden Tzach¹, Shiri Hodassman¹ and Ido Kanter¹,²*

¹Department of Physics, Bar-Ilan University, Ramat-Gan, 52900, Israel.
²Gonda Interdisciplinary Brain Research Center, Bar-Ilan University, Ramat-Gan, 52900, Israel.
*Corresponding author email: ido.kanter@biu.ac.il

This PDF file includes:

Figures. S1 and S2
Supplementary Figure S1. Probability density function (pdf) of $W^{\text{conv}}$ gradients, $|\Delta|$, and $|\Delta/W^{\text{conv}}|$ for Tree-3 using ReLU activation function. 

a, pdf of the absolute value of $|\Delta|$ for a test example of with a correct predicting label. The vertical red line stands for $\Delta_0$ (denoted in red) such that the summation of pdf($\Delta < \Delta_0$) ~0.97. 
b, Similar to (a) for $|\Delta/W^{\text{conv}}|$. 
c, Similar to (a) with a wrong predicting label. 
d, Similar to (b) with a wrong predicting label. Each one of the histograms consists of 1000 bins. In all panels, the vertical axis is in log-scale.
Supplementary Figure S2. Probability density function (pdf) of $W^{\text{conv}}$ gradients, $|\Delta|$, and $|\Delta/W^{\text{conv}}|$ for Tree-3 using Sigmoid activation function. a, pdf of the absolute value of $|\Delta|$ for a test example with a correct predicting label. The vertical red line stands for $\Delta_0$ (denoted in red) such that the summation of pdf($\Delta < \Delta_0$) ~0.97. b, Similar to (a) for $|\Delta/W^{\text{conv}}|$. c, Similar to (a) with a wrong predicting label. d, Similar to (b) with a wrong predicting label. Each one of the histograms consists of 1000 bins. In all panels, the vertical axis is in log-scale.