Activation Function Design for Deep Networks: Linearity and Effective Initialisation

Michael Murray

University of Oxford, Radcliffe Observatory
Oxford OX2 6GG, Great Britain

The activation function deployed in a deep neural network has great influence on the performance of the network at initialisation, which in turn has implications for training. In this talk we discuss how to avoid two problems at initialisation identified in prior works: in particular rapid convergence of pairwise input correlations, and vanishing and exploding gradients. The key takeaway from this talk will be that both these problems can provably be avoided by choosing an activation function possessing a sufficiently large linear region around the origin, relative to the bias variance of the network's random initialisation. Furthermore, we will observe this particular property appears key, with initialisation and training outcomes being fairly insensitive, within reason, to the particular shape of the activation function away from the origin. Beyond presenting theoretical results we will demonstrate empirically that using activation functions with this property leads to tangible benefits in practice, both in terms of test and training accuracy as well as training time.