Your 598 ID: 5107

Title of paper: Learning Functions: When is Deep Better than Shallow?

What is their primary result? The paper compares shallow (one hidden layer) networks with deep networks (in particular, the idealized model of a deep network as a binary tree) and shows that although both the shallow and deep neural networks are capable of achieving the same degree of accuracy the number of parameters, VC-dimension, and fat-shattering dimension are much smaller for the deep neural network.

Why is this important? The paper is a worthwhile crack at explaining why deep neural networks work better than shallow networks in practice.

What are their key ideas? Universality Theorem–if the activation function is analytic, but not polynomial anywhere. (Note that the usual ReLu, pReLu, etc., activation functions fail this), then any function that has r-continuous partial derivatives can be approximated with error $\mathrm{O}(n^{-r/d})$ by a sufficiently deep network. The authors prove that this is the best estimate that can be made given the set of assumptions.

Gaussian networks — the authors prove that if there's an approximating neural network with a Gaussian function that achieves the above approximation for a function, then the function has that many partial derivatives.

What are the limitations, either in performance or applicability? The result requires that the function we wish to approximate live in a smoothness space. There are more valuable smoothness spaces than the ones the author uses, so this work can be considered preliminary.

What might be an interesting next step based on this work? The work should be expanded to consider different function spaces.

What's the architecture?

How did they train and evaluate it?

Did they implement something?

Grader's 598 ID: