Basic Information

How to choose an activation function
Author: H. N. Mhaskar (CSLA), C. A. Micchelli (IBM. Watson) & Anne Schroeder
Proceeding NIPS'93 Proceedings of the 6th International Conference on Neural Information Processing Systems, pages 319-326
Citation: 39

Summary

This paper focuses on the complexity of estimating real-valued functions using neural networks. Complexity means approximating the number of neurons required to achieve a certain level of estimation accuracy. Previous work have proved that single hidden layer networks with any non-linear activation function can estimate any continuous function to any desired degree of accuracy. However, people still do not know how to construct (how many neurons to use) such networks given a known class of functions before this paper was published. Barron's theory only gave the results on sigmoidal activation function estimation for functions under certain conditions. Other work proved that functions with a bounded gradient with s variables require $O(n^s)$ neurons. In this paper, the authors focused on approximating functions with a certain number of smooth derivatives, which expand the functions that can be estimated. They investigate the number of neurons requires given activation functions. In addition to examining sigmoid activation function, they also consider more classes of functions. They gave a method to construct such networks with radial basis activation functions, including Gaussian, squashing and sigmoidal functions. They also showed that the smoother the activation function, the better the rate of approximation. This paper provides an overview for choosing activation functions and expands the deal-able functions to a larger class.

Strong points

1. **Originality.**
The idea is not new according to the related work stated by the authors. This paper only examines more classes of activation functions based on previous work. However, the idea of using Fourier coefficients and infinitely many times continuously differentiable function to estimate is interesting and useful.

2. **Technical detail.**
The paper is theoretically well-founded on the following points. The authors clearly stated the notations and background knowledge. The authors mentioned that this method is also valid for $L^p - norms$ approximation, instead of only being valid for uniform approximation.

3. **Writing**
I like the way the authors decompose the problem. They added conditions so that the problem can be reduced to a simplified version but without loss of generality. The conversion made the proof procedure simpler and easier for readers to understand.

Weak points

1. **Experiment/Example.**
This is a purely theoretical work. Although experiments may not be easy to implement for theoretical work, I think more concrete examples could be made. For Table 1, most results are remarkable. However, when I am reading the table, I care more about the comparison of the most common functions we are using today, including $ReLU(x^{k=1}$ if $x > 0$, 0, otherwise). However, the paper only considered $k > 1$), $sigmoid$, $tanh$ and so on. Unfortunately, famous activations such as $ReLU$ and $tanh$ are not examined in this table. Computing the result for a certain activation function requires much mathematics foundations and is not easy for readers to accomplish in a short time. Additionally, the authors also did not explicitly prove the relation between continuity and complexity to estimate, which is stated as a conclusion in the abstract.

2. **Technical detail.**
The paper mentioned some results with l, which is the number of hidden layers. However, the authors did not mention how to decide the number of neurons used in each layer. This makes the statements unclear.

3. **Writing**
Most core results have been published in the authors' previous work. They could have explained the most important conclusion instead of referring the reader to their other work. A brief explanation could help the readers better understand their results.