Basic Information

ResNet with one-neuron hidden layers is a Universal Approximator.
Author: Hongzhou Lin, Stefanie Jegelka (MIT)
Nips 18, Cited by 7

Summary

This paper demonstrate a deep ResNet can uniformly approximate any Lebesgue integrable function in d dimensions. Current success in machine learning is largely due to the development of deep neural networks. Deeper networks mean better representation ability. But can deep neural networks have the ability to approximate any function is an interesting but unsolved problem. Work in the late eighties, the universal approximation theorem showed single hidden layer networks(but may contain infinite neurons) can approximate any continuous function with compact support.

The main contribution of this paper is to show that depth can also lead to universal approximation and more efficient than the wide networks. It proves that the ResNet with one single neuron per hidden layer is enough to approximate any Lebesgue integrable function: $\forall f : \mathbb{R}^d \to \mathbb{R}, \forall \epsilon > 0, \exists R, s.t. \int_{\mathbb{R}^d} |f(x) - R(x)|dx \leq \epsilon$ where R is a ResNet with ReLU activation function and one neuron per hidden layer.

They carried out an experiment of a unit ball to compare narrow fully connected networks with ResNet. They showed that d is too narrow for fully connected networks to achieve universal approximation for they are unable to approximate unbounded region. They further prove that ResNet can achieve universal approximation by giving a constructive solution to approximate piecewise constant functions. They give a draft proof for $d = 1$. Since piecewise constant functions with compact support and finitely many discontinuities is dense in $l_1(\mathbb{R}^d)$, such deep ResNets are also dense in $l_1(\mathbb{R}^d)$. They also give details about how to adjust the construction to keep the function unchanged on previous subdivisions. Then they extend the problem to higher dimension.

At last, they give several technical analysis of the hidden units/layers number. They also mentioned that the training efficiency is not guaranteed. Though generalization ability is not proved but works well in practice.

Strong points

1. **Originality.**
This paper gives an interesting way of prove the approximation ability of ResNet. The main difficult is that ResNet architecture does not allow plus operation of different functions. So maintaining the information in the previous layer is a huge concern. They successfully give a construction method for approximation.
The idea of prove fully connected networks are not able to approximate because they cannot approximate bounded region is also novel and interesting.

2. **Writing**
They organized the paper in a really good way. Introducing this problem with the unit ball example helps me to understand the motivation. The proof also goes fluently and easy to understand. They also add clear figures to help understand the construction and adjustment of the functions.

3. **Soundness**
I think it is good for the authors to cover the training effectiveness and the generalization issue. These are two main consideration in practice. These statements make the paper sound and solid. It also provides a future research direction.

4. **Contribution**
This paper gives a proof of the effectiveness of the ResNet. This structure allows networks to go deeper and achieve better representation. This provides a theoretical support for the ResNet and enables the further development in deep learning.

Weak points

1. **Experiment**
Though they mentioned that the training efficiency is not guaranteed generally, I think they could have carry out several experiment to compare the time complexity of fully connect networks and the ResNets.
This is also true for the generalization conclusion. The motivation example is just fake data. Fully connected networks may fail on a certain distribution but so could the ResNet.

2. **Technical detail.**
When one prove the existent, they could restrain the number of neurons so the prediction can achieve a certain accuracy. However, I believe the reason to restrain the width of each layer to d is not clearly states in the paper. It is possible that fully connected networks could work well on a slightly wider network.