Beyond Finite Layer Neural Network: Bridging Deep Architects and Numerical Differential Equations

Yiping Lu, Aoxiao Zhong, QuanZheng Li and Bin Dong

Introduction

In this work, we show that many effective networks, such as ResNet, PolyNet, FractalNet and RevNet, can be interpreted as different numerical discretizations of differential equations。
We also propose a linear multi-step architecture (LM-architecture) which is inspired by the linear multi-step method solving ordinary differential equations. The LM-architecture is an effective structure that can be used on any ResNet-like networks。

Neural Network As Numerical Scheme

Our motivation is to consider the residual network $x_n = x_{n-1} + f(x_{n-1})$ as the forward Euler scheme for the ODE $\dot{X} = f(X)$. We found out that many neural network can be consider as different numerical scheme for ODE , some examples is listed below.

Table1. Neural Networks' associated ODES and numerical scheme

Network	Related ODE	Numerical Scheme
ResNet, ResNeXt	$u_t = f(u)$	Forward Euler
PolyNet	$u_t = f(u)$	Approximation of Backward Euler
Fractal Net	$u_t = f(u)$	Runge-Kutta
RevNet	$u_t = f(v), v_t = g(u)$	Forward Euler

LM(Linear Multistep)-ResNet

We utilize the linear multi-step scheme to discrete the ODE, then we get a new type of neural network:

$$x_n = (1 - k_n)x_{n-1} + k_n x_{n-2} + f(x_{n-1})$$

Our neural network only inject one parameter in each layer of a ResNet but shows a huge improvement.

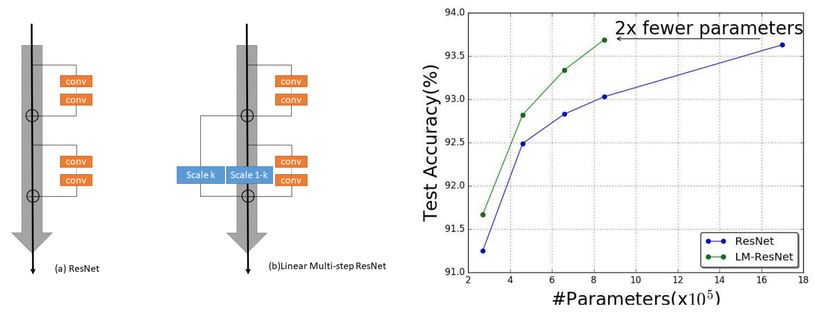


Fig1. Our LM-ResNet and numerical result on CIFAR10.

Benefit Of LM Structure

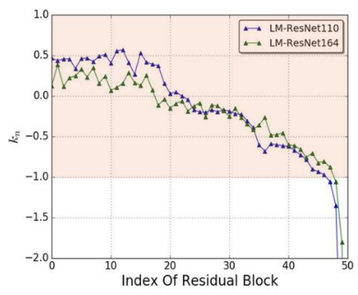
We analysis our scheme by modified equation, which is a new equation approximated by the scheme in a higher order.

ResNet:

$$\frac{\Delta t}{2}\ddot{X} + \dot{X} = f(u)$$

LM-ResNet:

$$(1 - k_n)\frac{\Delta t}{2}\ddot{X} + (1 + k_n)\dot{X} = f(u)$$



In short, we find out that the parameterk_n learned in the LM-ResNet introduced a momentum in to the dynamic of information propagating in network.

Table2. LM-ResNeXt On CIFAR100

Model	Layer	Top1	Top5
ResNeXt	29(8x64d)	17.77	34.4M
ResNeXt	29(16x64d)	17.31	68.1M
ResNeXt	29(16x64d), pre-act	17.65	68.1M
LM-ResNeXt	29(8x64d), pre-act	17.49	34.4M
LM-ResNeXt	29(16x64d), pre-act	16.79	68.1M

Table3. Single-crop Error Rate On Imagenet

Model	Layer	Top1	Top5
ResNet	50	24.7	7.8
ResNet	101	23.6	7.1
ResNet	153	23.0	6.7
LM-ResNet	50, pre-act	24.0	7.3
LM-ResNet	101. pre-act	22.6	6.4

Stochastic Learning Meets Stochastic Control

We also find some stochastic learning method can be consider as approximating a stochastic dynamic system.

Stochastic Depth

$$dX = p(t)f(X)dt + \sqrt{p(t)(1-p(t))}f(X) \odot [1_{N\times 1}, 0_{N\times 1}]dB_t$$

Shake-Shake

$$dX = \frac{1}{2}(f_1(X) + f_2(X))dt + \frac{1}{\sqrt{12}}(f_1(X) - f_2(X)) \odot [1_{N\times 1}, 0_{N\times 1}]dB_t$$

So that we can consider stochastic learning method as solving a general stochastic control problem.

$$\min \mathbb{E}_{X(0)\sim data}\left(\mathbb{E}\left(l(X(t))\right) + \int_0^t R(\theta)\right)$$
$$s.t. \partial_t X = f(X,\theta) + g(X,\theta)dB_t$$

We also tested stochastic learning version of LM-ResNet, which also boosted the performance.

Model	Layer	Training Strategy	Error
ResNet	110		6.61
ResNet	110,pre-act		6.37
ResNet	110	Stochastic Depth	5.25
ResNet	1202	Stochastic Depth	4.91
LM-ResNet	56,pre-act	Stochastic Depth	5.14
LM-ResNet	110,pre-act	Stochastic Depth	4.80

Table4. Result Of LM-ResNet With Stochastic Depth On CIFAR10

Reference

E, Weinan. A proposal on machine learning via dynamical systems. Communications in Mathematics and Statistics, 5(1):1–11, 2017.
He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Identity mappings in deep residual networks. IEEE Conference on Computer Vision and Pattern Recognition, 2016.