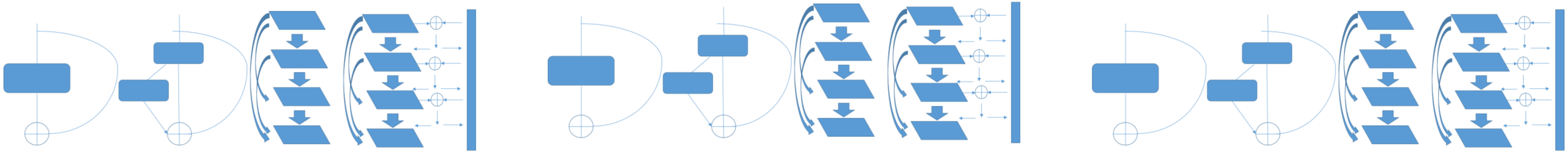
Beyond finite layer neural network

Bridging Numerical Dynamic System And Deep Neural Networks

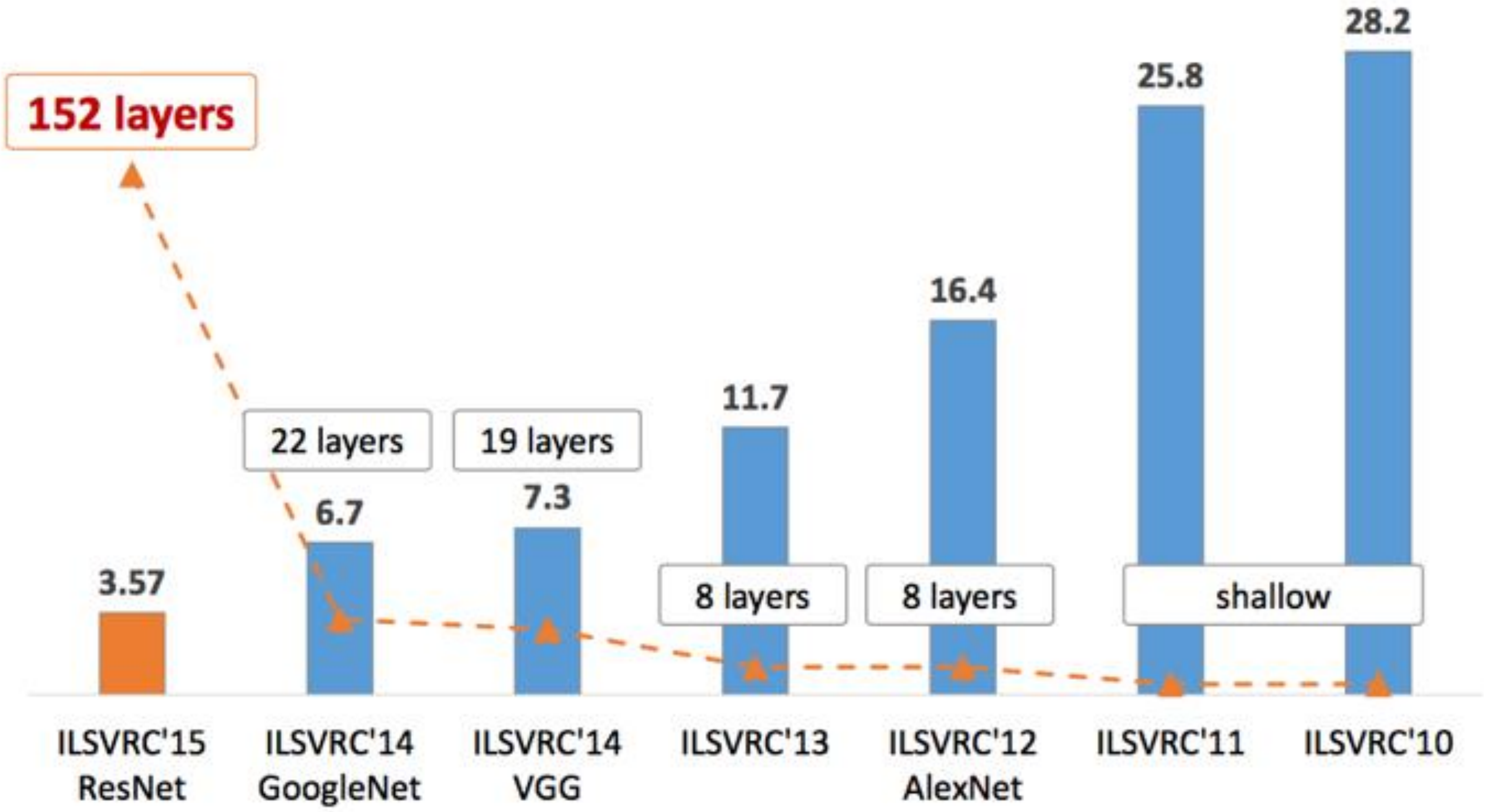
arXiv:1710.10121

Joint work with Bin Dong, Quanzheng Li, Aoxiao Zhong

Yiping Lu
Peiking University
School Of Mathematical Science

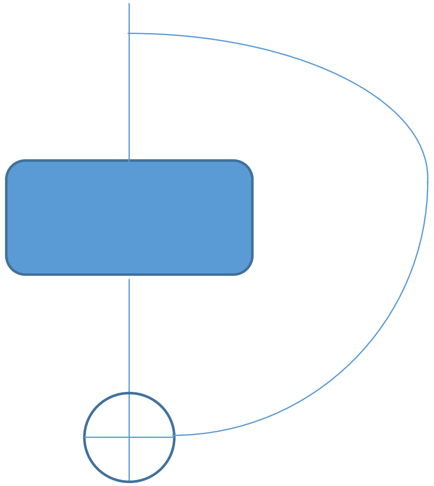


Depth Revolution



Motivation

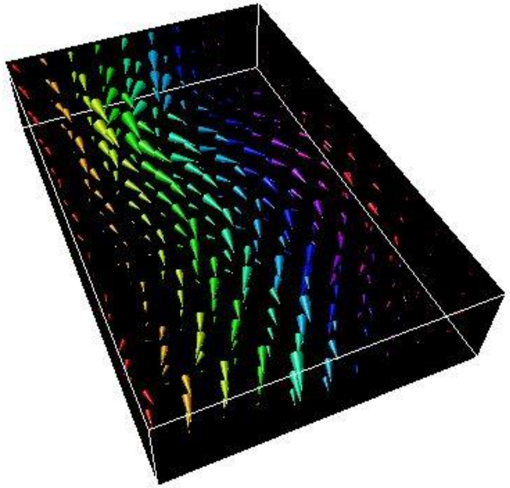
Deep Residual Learning(@CVPR2016)



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

Forward Euler Scheme

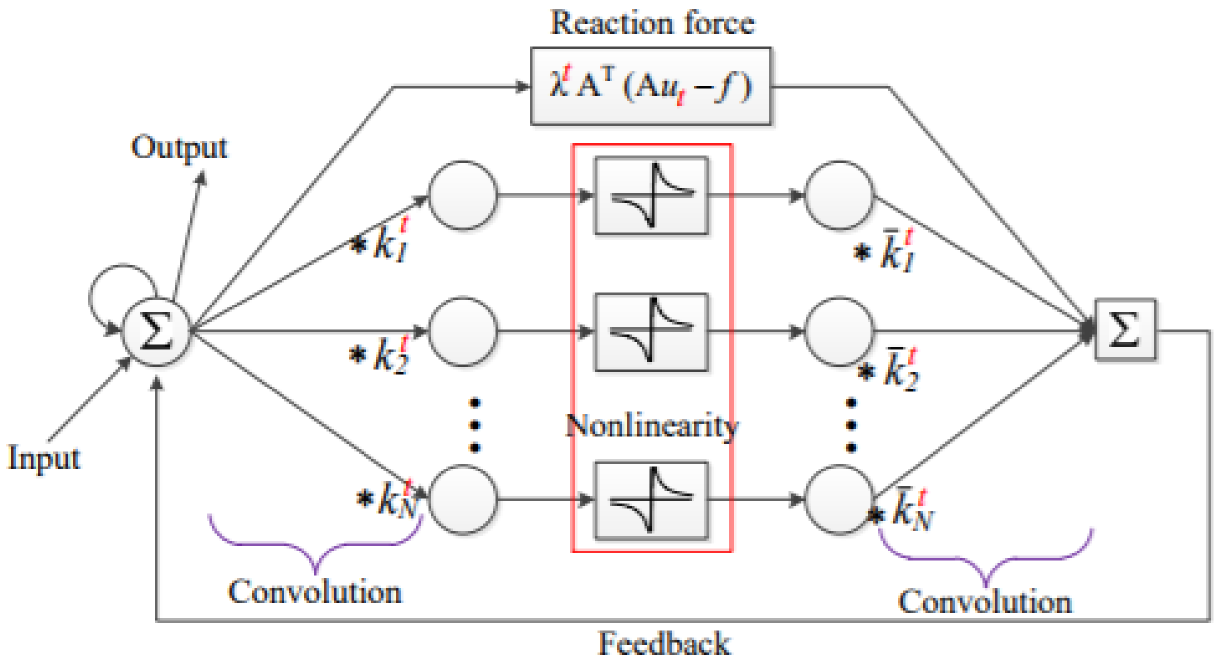
$$x_t = f(x)$$



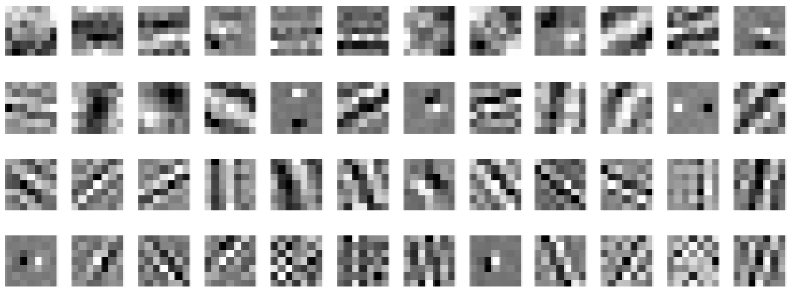
Weinan E. A Proposal on Machine Learning via Dynamical Systems.

Previous Works

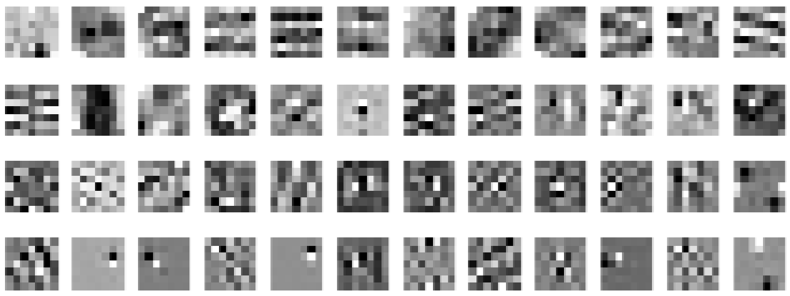
TRD(@CVPR2015): learn a diffusion process for denoising



Method	256^2	512^2	1024^2	2048^2	3072^2
BM3D [10]	1.1	4.0	17	76.4	176.0
$\mathrm{CSF}^5_{7\times7}$ [37]	3.27	11.6	40.82	151.2	494.8
WNNM [18]	122.9	532.9	2094.6	–	–
$\mathrm{TRD}^5_{5\times5}$	0.51	1.53	5.48	24.97	53.3
	0.43	0.78	2.25	8.01	21.6
	0.005	0.015	0.054	0.18	0.39
$\mathrm{TRD}^5_{7\times7}$	1.21	3.72	14.0	62.2	135.9
	0.56	1.17	3.64	13.01	30.1
	0.01	0.032	0.116	0.40	0.87



(a) 48 filters of size 7 × 7 in stage 1



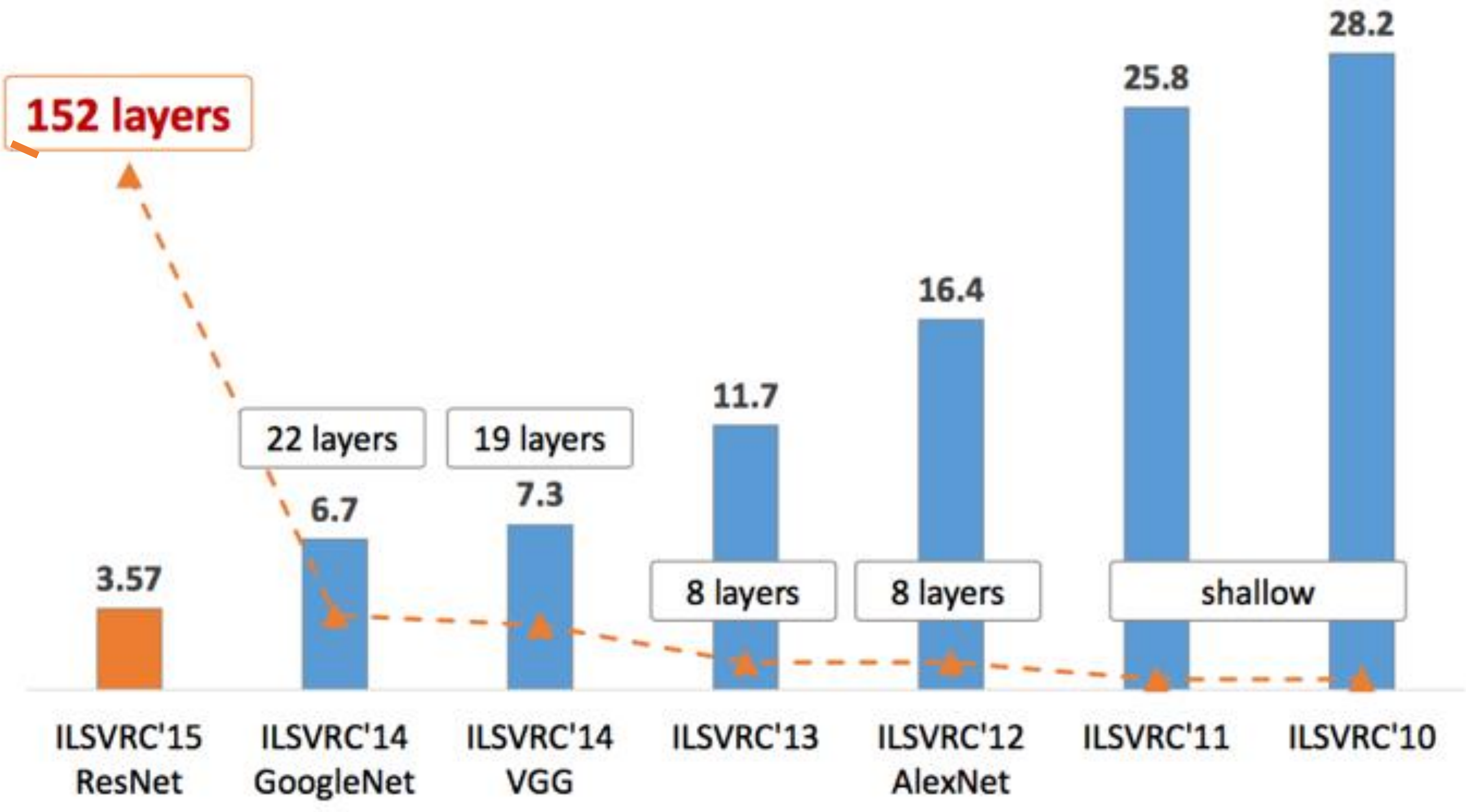
(b) 48 filters of size 7 × 7 in stage 5

Chen Y, Yu W, Pock T. On learning optimized reaction diffusion processes for effective image restoration CVPR2015

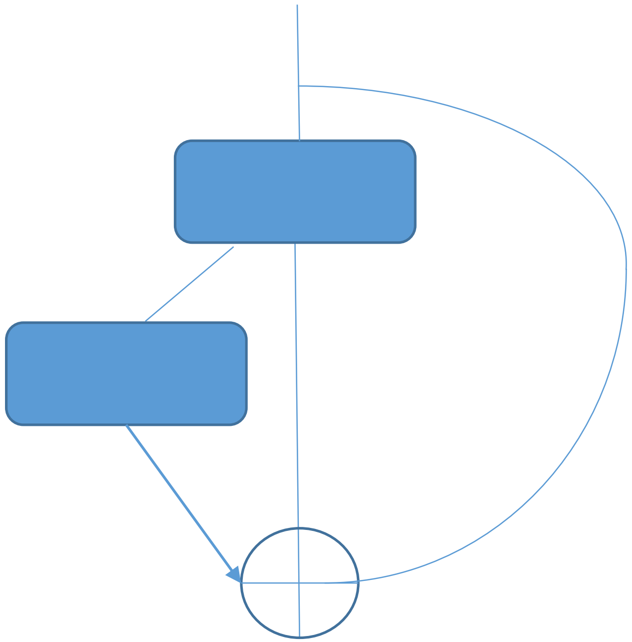
Depth Revolution

Going into infinite layer

Differential Equation
As Infinite Layer
Neural Network



Polynet(@CVPR2017)



(b) Polynet

Revisiting previous efforts in deep learning, we found that diversity, another aspect in network design that is relatively less explored, also plays a significant role

PolyStrure: $x_{n+1} = x_n + F(x_n) + F(F(x_n))$

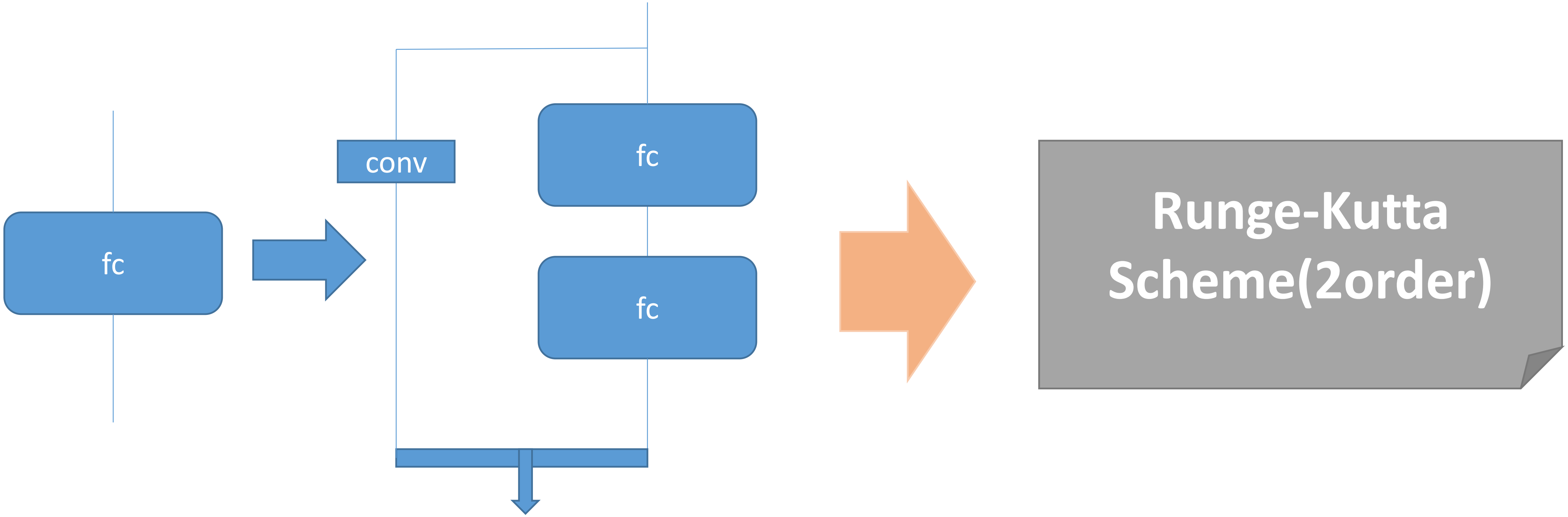
Backward Euler Scheme:

$$x_{n+1} = x_n + F(x_{n+1}) \Rightarrow x_{n+1} = (I - F)^{-1} x_n$$

Approximate the operator $(I - F)^{-1}$ by $I + F + F^2 + \cdots$

Zhang X, Li Z, Loy C C, et al. PolyNet: A Pursuit of Structural Diversity in Very Deep Networks

FractalNet(@ICLR2017)



$$x_{n+1} = k_1 x_n + k_2(k_3 x_n + f_1(x_n)) + f_2(k_3 x_n + f_1(x_n))$$

Larsson G, Maire M, Shakhnarovich G. FractalNet: Ultra-Deep Neural Networks without Residuals.

PDE: Infinite Layer Neural Network

Dynamic System — **Nueral Network**

Continuous limit — Numerical Approximation

Table 1: In this table, we list a few popular deep networks, their associated ODEs and the numerical schemes that are connected to the architecture of the networks.

Network	Related ODE	Numerical Scheme
ResNet, ResNeXt, etc.	$u_t = f(u)$	Forward Euler scheme
PolyNet	$u_t = f(u)$	Approximation of backward Euler scheme
FractalNet	$u_t = f(u)$	Runge-Kutta scheme
RevNet	$\dot{X} = f_1(Y), \dot{Y} = f_2(X)$	Forward Euler scheme

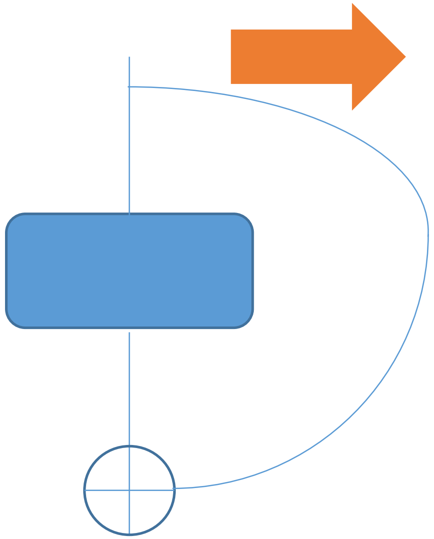
WRN, ResNeXt, Inception-ResNet, PolyNet, SENet etc...... :

New scheme to Approximate the right hand side term

Why not change the way to discrete u_t?

Experiment

@Linear Multi-step Residual Network

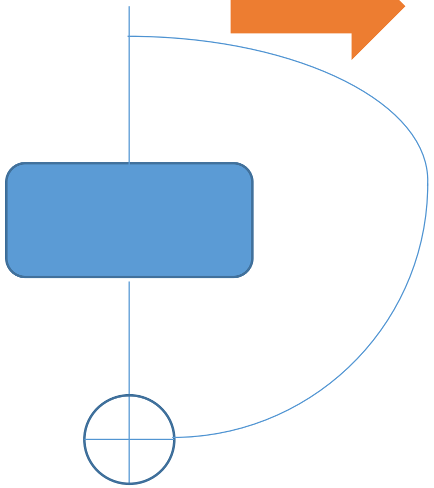


$$x_t = f(x)$$

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

Experiment

@Linear Multi-step Residual Network

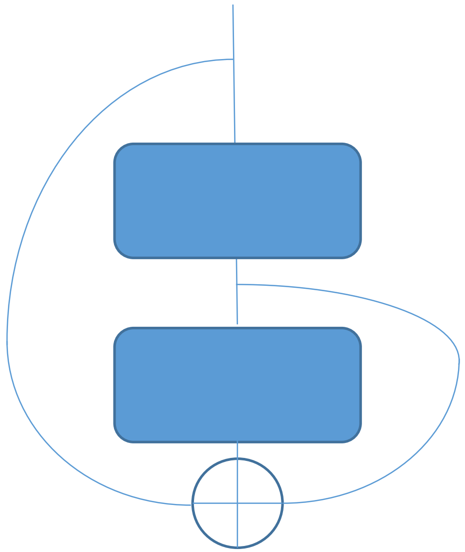


$$x_{n+1} = x_n + f(x_n)$$

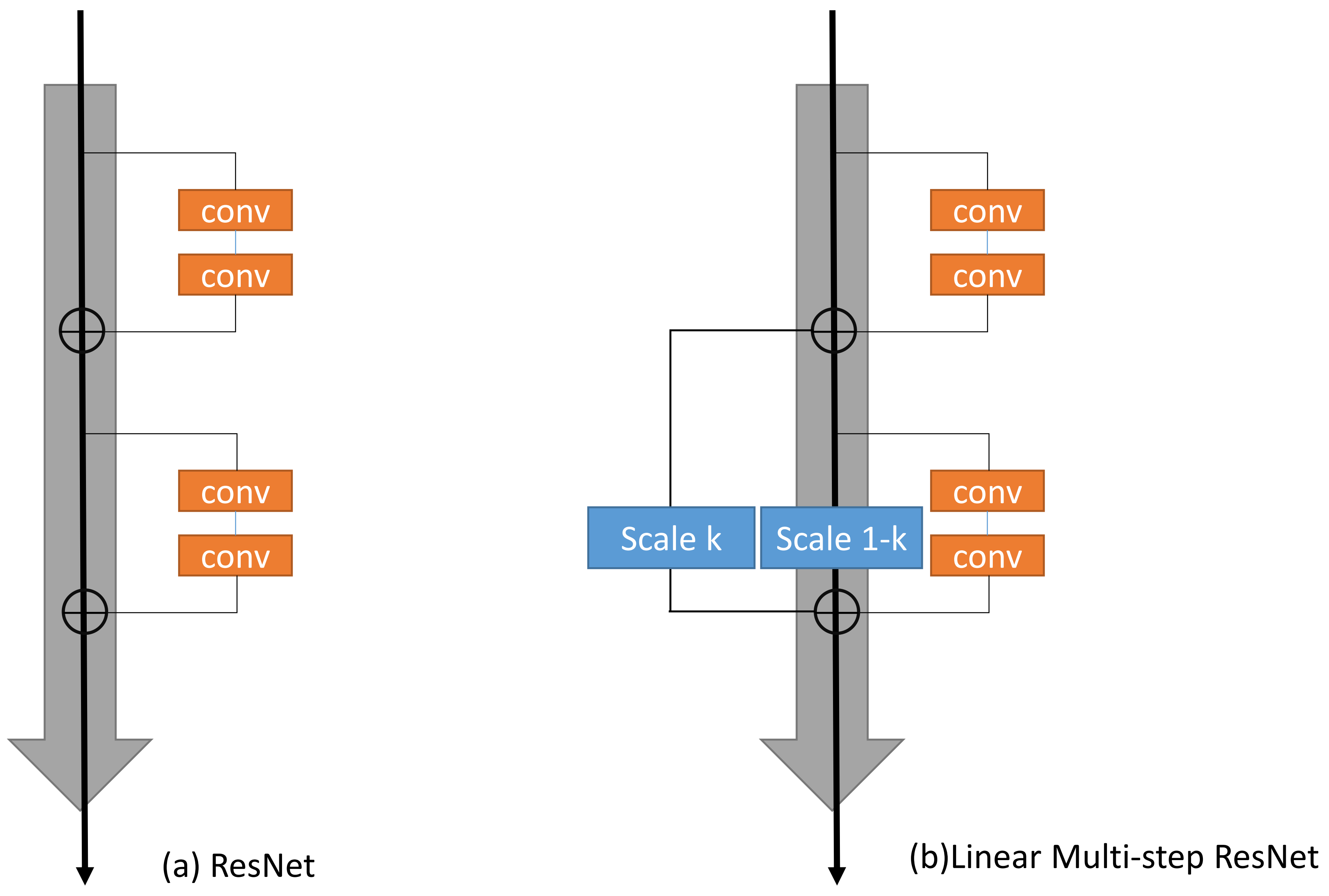
$$x_t = f(x)$$

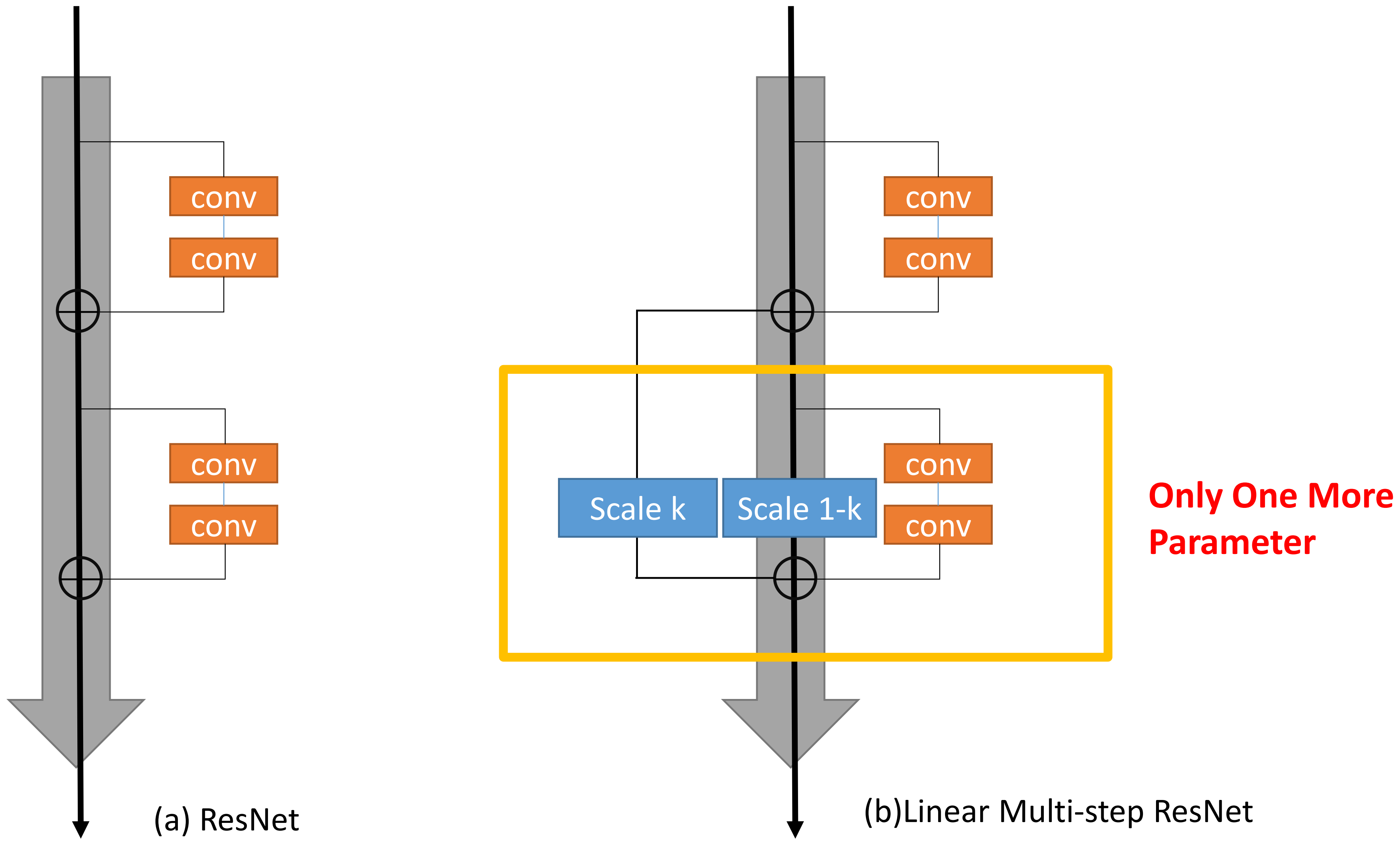
Linear Multi-step Scheme

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



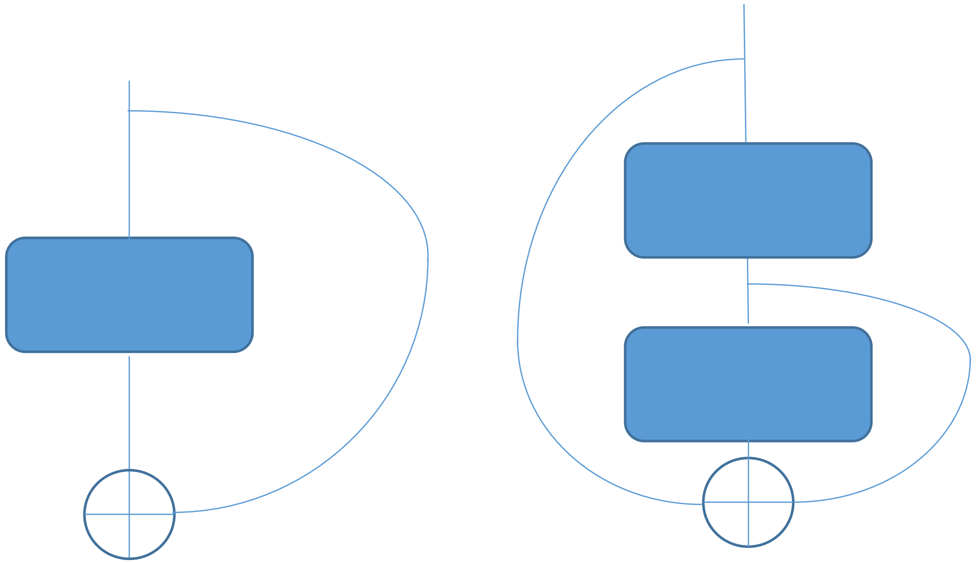
Linear Multi-step Residual Network


conv
conv
conv
conv
(a) ResNet
conv
conv
Scale k
Scale 1-k
conv
conv
(b)Linear Multi-step ResNet


conv
conv
conv
conv
(a) ResNet
conv
conv
Scale k
Scale 1-k
conv
conv
Only One More Parameter
(b)Linear Multi-step ResNet

Experiment

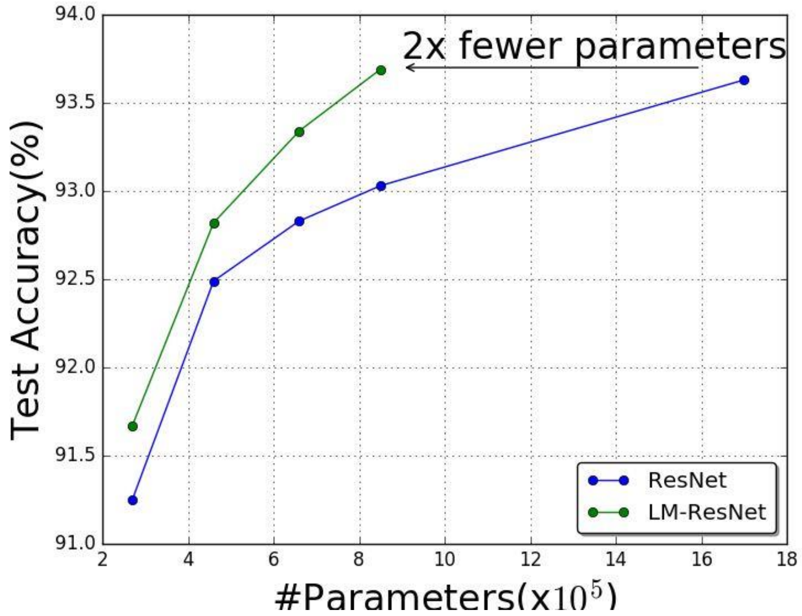
@Linear Multi-step Residual Network



(a)Resnet (b)LM-Resnet

Table 2: Comparisons of LM-ResNet/LM-ResNeXt with other networks on CIFAR

Model	Layer	Error	Params	Dataset
ResNet (He et al. (2015b))	20	8.75	0.27M	CIFAR10
ResNet (He et al. (2015b))	32	7.51	0.46M	CIFAR10
ResNet (He et al. (2015b))	44	7.17	0.66M	CIFAR10
ResNet (He et al. (2015b))	56	6.97	0.85M	CIFAR10
ResNet (He et al. (2016))	110, pre-act	6.37	1.7M	CIFAR10
LM-ResNet (Ours)	20, pre-act	8.33	0.27M	CIFAR10
LM-ResNet (Ours)	32, pre-act	7.18	0.46M	CIFAR10
LM-ResNet (Ours)	44, pre-act	6.66	0.66M	CIFAR10
LM-ResNet (Ours)	56, pre-act	**6.31**	0.85M	CIFAR10



Experiment

@Linear Multi-step Residual Network

Table 2: Linear Multi-step Resnet Test On Cifar

Model	Layer	Accuracy	Params	Dataset
Resnet	20	91.25	0.27M	Cifar10
Resnet	32	92.49	0.46M	Cifar10
Resnet	44	92.83	0.66M	Cifar10
Resnet	56	93.03	0.85M	Cifar10
Resnet	110	93.63	1.7M	Cifar10
LM-Resnet(Ours)	20	91.67	0.27M	Cifar10
LM- Resnet(Ours)	32	92.82	0.46M	Cifar10
LM- Resnet(Ours)	44	92.98	0.66M	Cifar10
LM- Resnet(Ours)	56	**93.69**	0.85M	Cifar10
EM- Resnet(Ours)	40	91.75	0.27M	Cifar10
Resnet	110	72.24	1.7M	Cifar100
Resnet	164	75.67	2.55M	Cifar100
Resnet	1202	77.29	18.88M	Cifar100
ResneXt	29(8×64d)	82.23	34.4M	Cifar100
ResneXt	29(16×64d)	82.69	68.1M	Cifar100
LM-Resnet(Ours)	110	73.16	1.7M	Cifar100
LM-Resnet(Ours)	164	76.74	2.55M	Cifar100
LM-ResneXt(Ours)	29(8×64d)	82.51	34.4M	Cifar100
LM-ResneXt(Ours)	29(16×64d)	**83.21**	68.1M	Cifar100

Table 3: Single-crop error rate on ImageNet (validation set)

Model	Layer	top-1	top-5
ResNet (He et al. (2015b))	50	24.7	7.8
ResNet (He et al. (2015b))	101	23.6	7.1
ResNet (He et al. (2015b))	152	23.0	6.7
LM-ResNet (Ours)	50, pre-act	23.8	7.0
LM-ResNet (Ours)	101, pre-act	**22.6**	**6.4**

Explanation on the performance boost via *modified equations*

@Linear Multi-step Residual Network

ResNet

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \Delta t \boldsymbol{f}(\boldsymbol{x}_n) \quad \Longrightarrow \quad \dot{\boldsymbol{u}} + \frac{\Delta t}{2} \ddot{\boldsymbol{u}}_n = \boldsymbol{f}(\boldsymbol{u})$$

LM-ResNet

$$\boldsymbol{x}_{n+1} = (1 - k_n)\boldsymbol{x}_n + k_n \boldsymbol{x}_{n-1} + \Delta t \boldsymbol{f}(\boldsymbol{x}_n) \quad \Longrightarrow \quad (1 + k_n)\dot{\boldsymbol{u}} + (1 - k_n)\frac{\Delta t}{2} \ddot{\boldsymbol{u}}_n = \boldsymbol{f}(\boldsymbol{u})$$

[1] Dong B, Jiang Q, Shen Z. Image restoration: wavelet frame shrinkage, nonlinear evolution PDEs, and beyond. Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal 2017.
[2] Su W, Boyd S, Candes E J. A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights. Advances in Neural Information Processing Systems, 2015.
[3] A. Wibisono, A. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimizationProceedings of the National Academy of Sciences 2016.

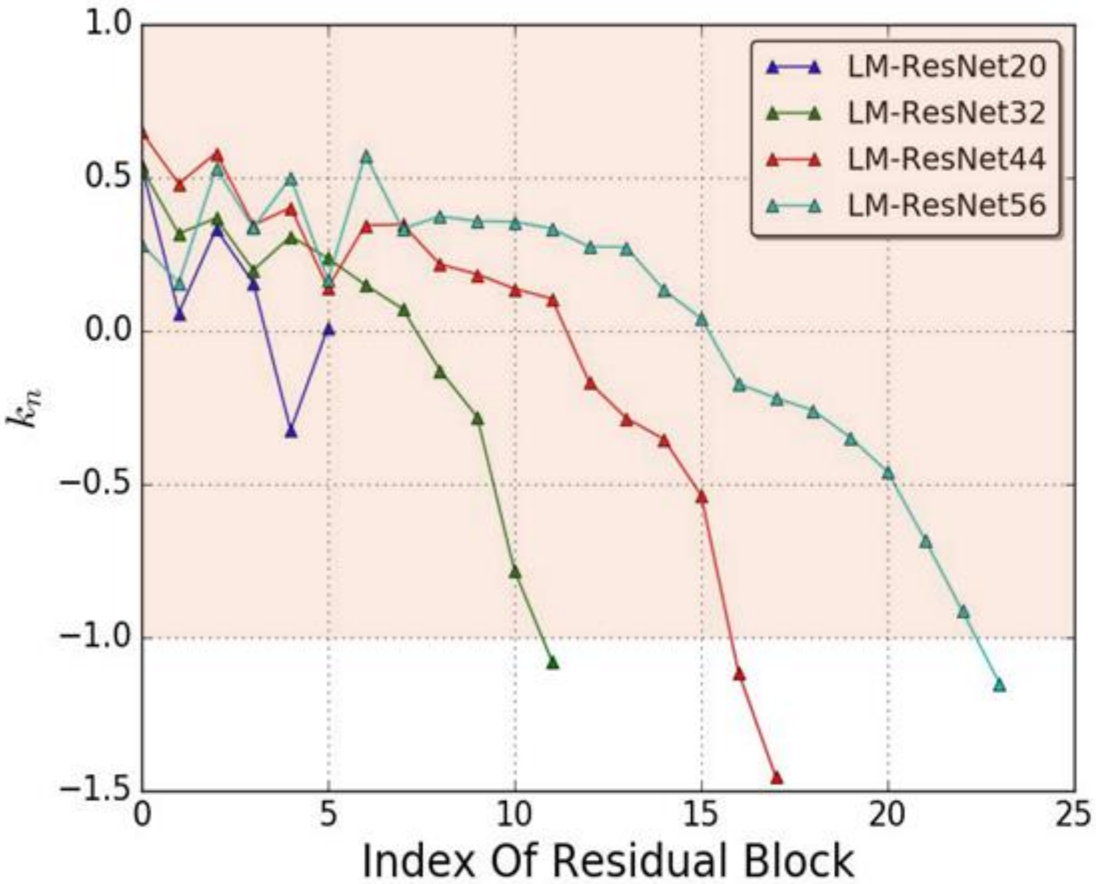
Plot The Momentum

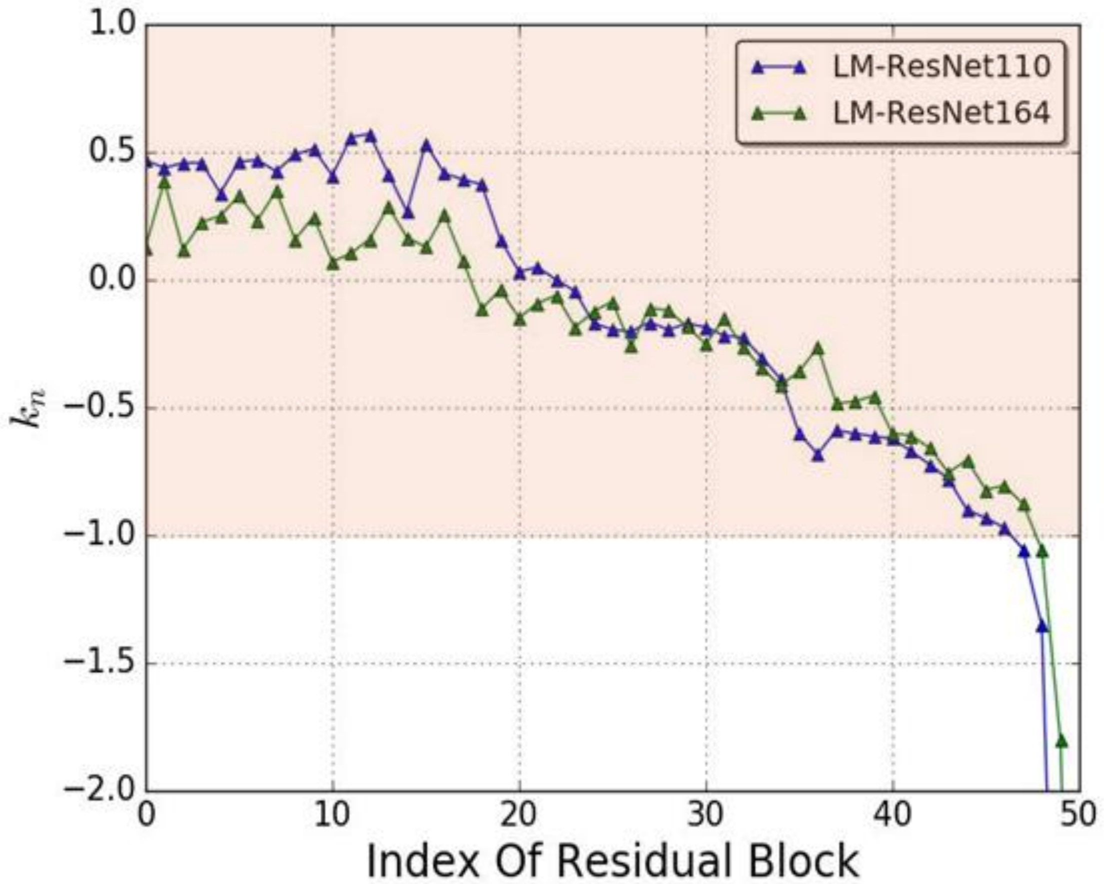
@Linear Multi-step Residual Network

$$\boldsymbol{x_{n+1} = (1 - k_n)x_n + k_n x_{n-1} + \Delta t f(x_n)}$$

Learn A Momentum

$$\boldsymbol{(1 + k_n)\,\dot{u} + \boxed{(1 - k_n)\frac{\Delta t}{2}\ddot{u}_n} + o(\Delta t^3) = f(u)}$$





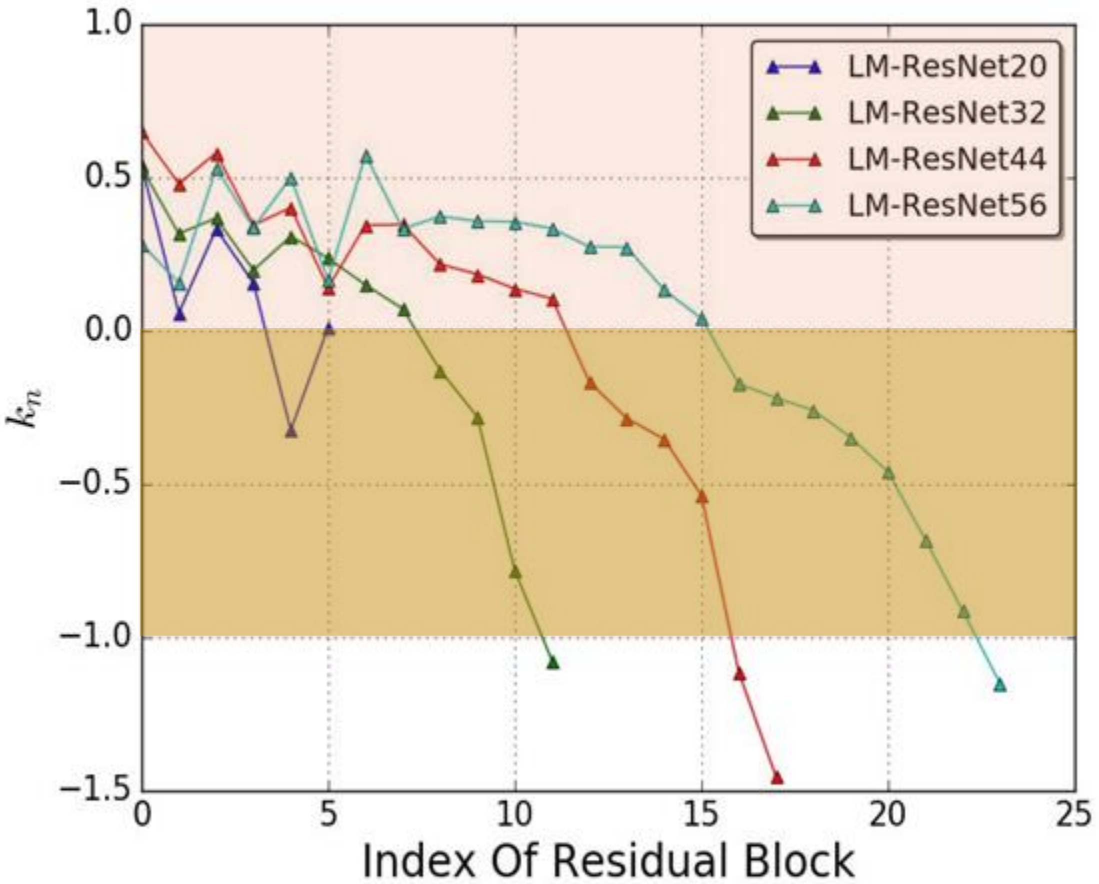
Plot The Momentum

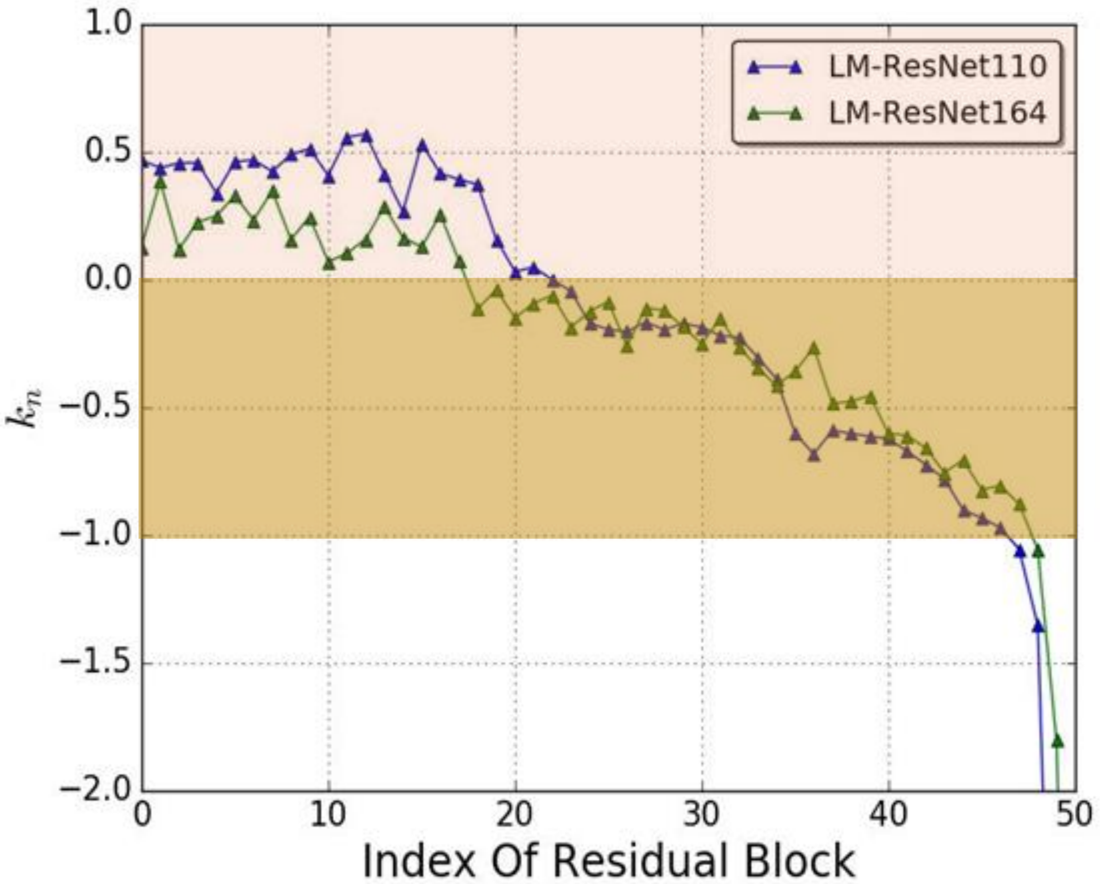
@Linear Multi-step Residual Network

$$x_{n+1} = (1 - k_n)x_n + k_n x_{n-1} + \Delta t f(x_n)$$

Learn A Momentum

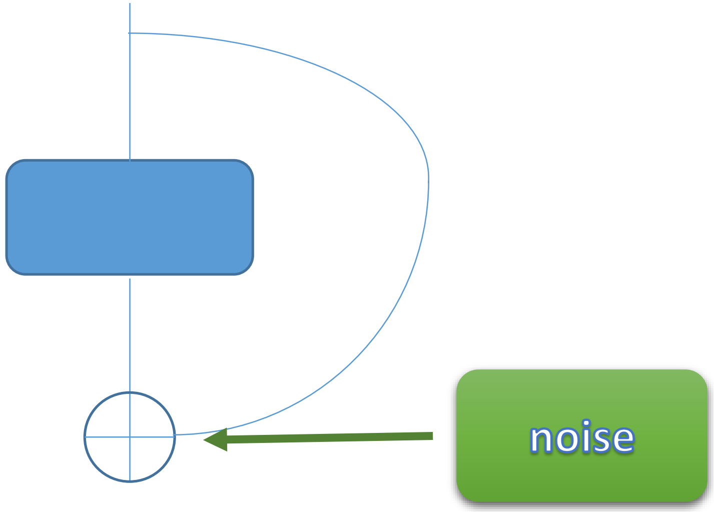
$$(1 + k_n)\,\dot{u} + \boxed{(1 - k_n)\frac{\Delta t}{2}\ddot{u}_n} + o(\Delta t^3) = f(u)$$



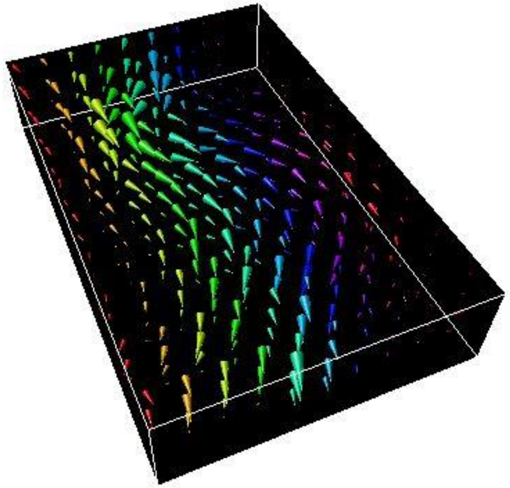


Bridge the stochastic dynamic

Noise can avoid overfit?



Dynamic System



Previous Works

Shake-Shake regularization

$$x_{n+1} = x_n + \eta f_1(x) + (1-\eta) f_2(x), \eta \sim U[0,1]$$
$$= x_n + f_2(x_n) + \frac{1}{2}\left(f_1(x_n) - f_2(x_n)\right) + \boxed{(\eta - \frac{1}{2})\left(f_1(x_n) - f_2(x_n)\right)}$$

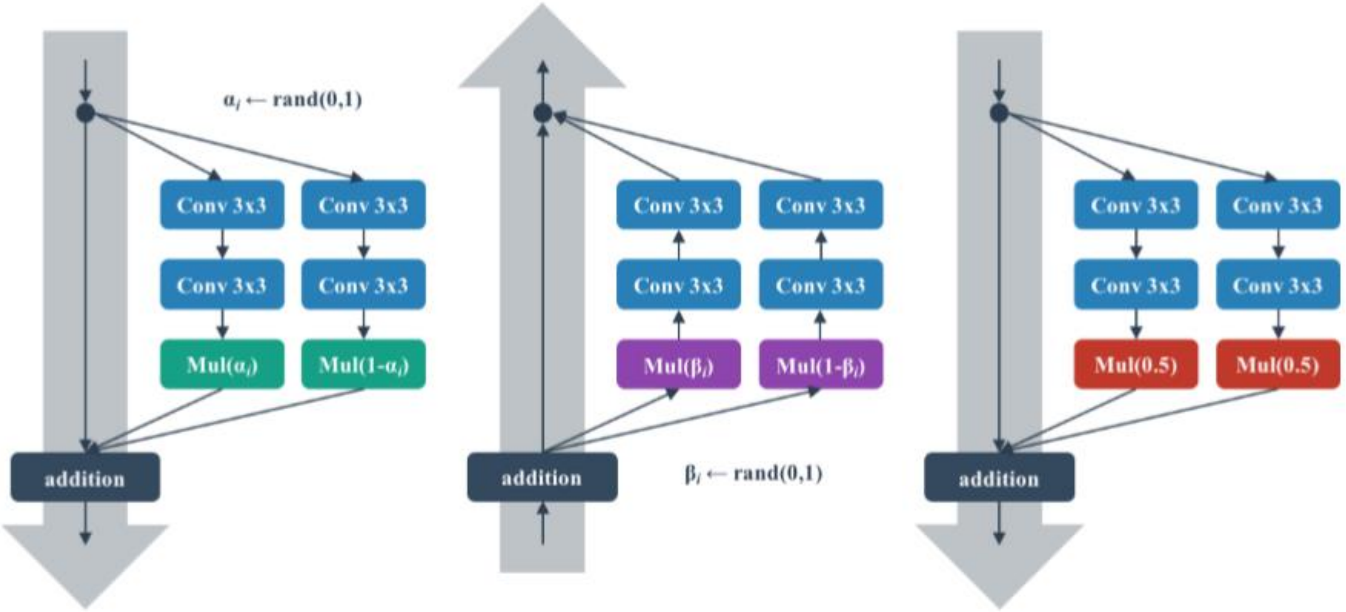


Figure 1: **Left:** Forward training pass. **Center:** Backward training pass. **Right:** At test time.

$$\frac{1}{\sqrt{12}}(f_1(X) - f_2(X)) \odot [\mathbf{1}_{N\times 1}, \mathbf{0}_{N,N-1}]dB_t$$

Apply data augmentation techniques to internal representations.

Gastaldi X. Shake-Shake regularization. ICLR Workshop Track2017.

Previous Works

Deep Networks with Stochastic Depth

$$x_{n+1} = x_n + \eta_n f(x)$$
$$= x_n + E\eta_n f(x_n) + \boxed{(\eta_n - E\eta_n) f(x_n)}$$

$$\sqrt{p(t)(1-p(t))} f(X) \odot [\mathbf{1}_{N\times 1}, \mathbf{0}_{N,N-1}] dB_t.$$

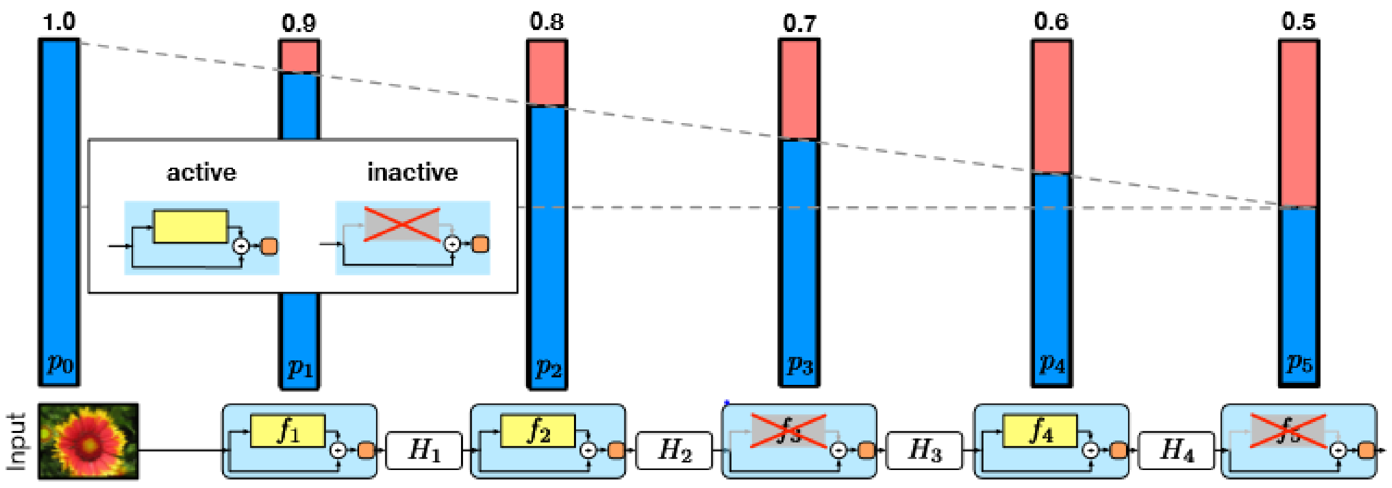


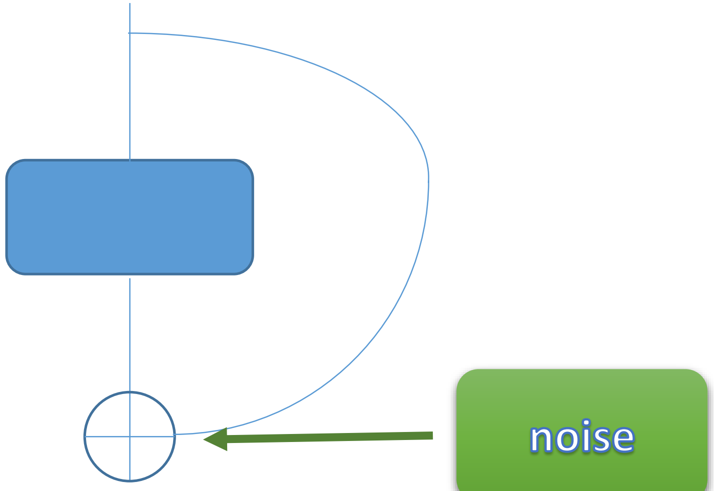
Fig. 2. The linear decay of p_ℓ illustrated on a ResNet with stochastic depth for $p_0 = 1$ and $p_L = 0.5$. Conceptually, we treat the input to the first ResBlock as H_0, which is always active.

To reduce the effective length of a neural network during training, we randomly skip layers entirely.

Huang G, Sun Y, Liu Z, et al. Deep Networks with Stochastic Depth ECCV2016.

Bridge the stochastic control

Noise can avoid overfit?



$$\dot{X}(t) = f\big(X(t), a(t)\big) + g(X(t), t)dB_t, X(0) = X_0$$

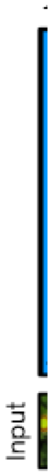
The numerical scheme is only need to be **<u>weak convergence</u>**!

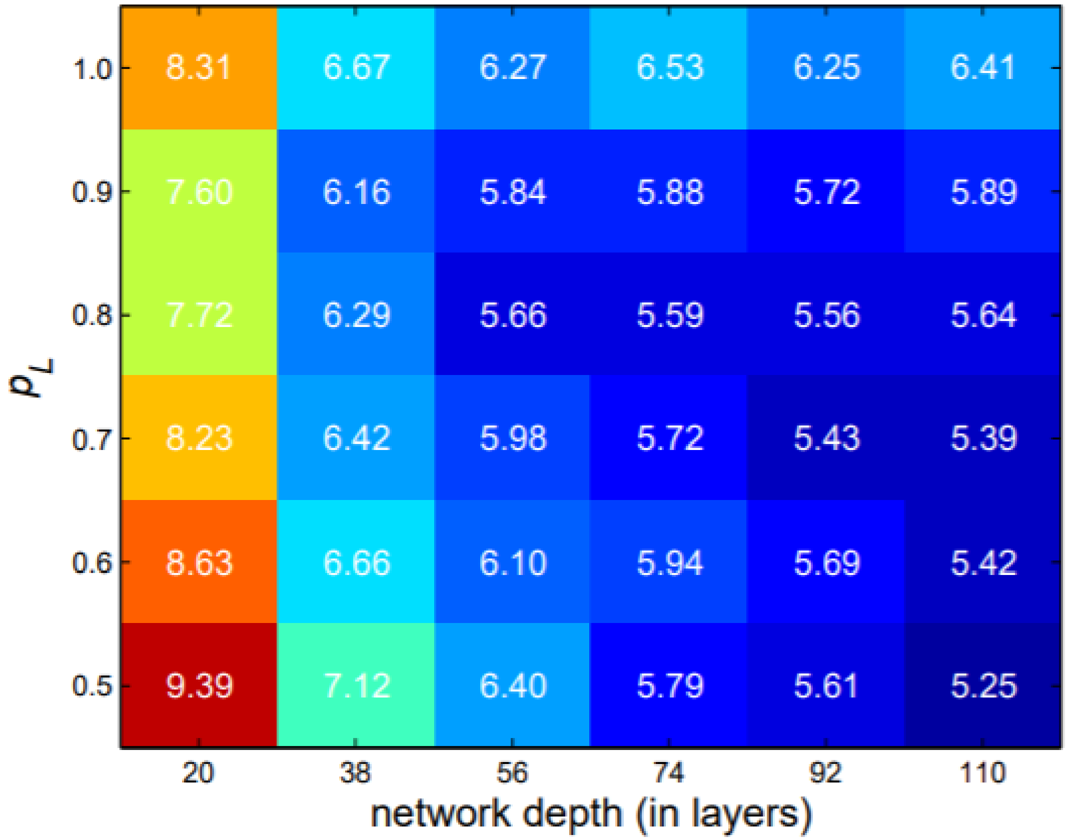
Previous Works

Deep Networks with Stochastic Depth

$$x_{n+1} = x_n + \eta_n f(x)$$
$$= x_n + E\eta_n f(x_n) + \boxed{(\eta_n - E\eta_n) f(x_n)}$$

We need $1 - 2p_n = O(\sqrt{\Delta t})$





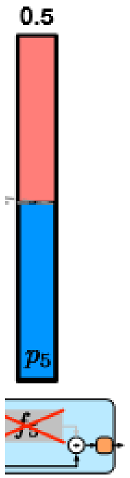


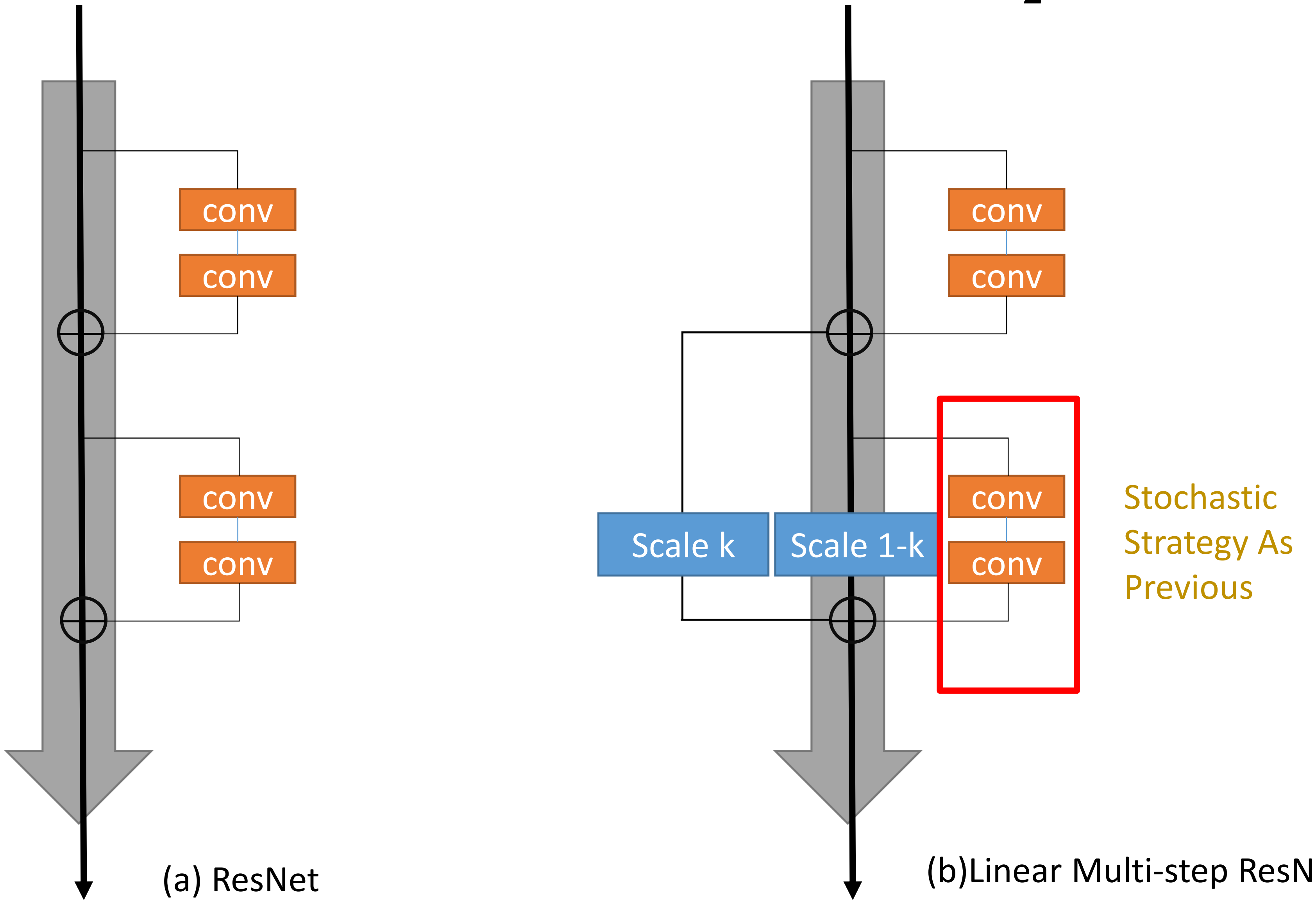
Fig
and
alw

or $p_0 = 1$
which is

To reduce the effective length of a neural network during training, we randomly skip layers entirely.

Huang G, Sun Y, Liu Z, et al. Deep Networks with Stochastic Depth ECCV2016.

$$\left(\mathbf{1}+k_n\right)\dot{u}+\left(\mathbf{1}-k_n\right)\frac{\Delta t}{2}\ddot{u}_n+o\left(\Delta t^3\right)=f(u)+g(u)dW_t$$



(a) ResNet

(b)Linear Multi-step ResNet

Experiment

@Linear Multi-step Residual Network

Table 4: Test on stochastic training strategy on CIFAR10

Model	Layer	Training Strategy	Error
ResNet(He et al. (2015b))	110	Original	6.61
ResNet(He et al. (2016))	110,pre-act	Orignial	6.37
ResNet(Huang et al. (2016b))	56	Stochastic depth	5.66
ResNet(Our Implement)	56,pre-act	Stochastic depth	5.55
ResNet(Huang et al. (2016b))	110	Stochastic depth	5.25
ResNet(Huang et al. (2016b))	1202	Stochastic depth	4.91
ResNet(Ours)	110,pre-act	Gaussian noise (noise level = 0.001)	5.52
LM-ResNet(Ours)	56,pre-act	Stochastic depth	5.14
LM-ResNet(Ours)	110,pre-act	Stochastic depth	**4.80**

Conclusion

@Beyond Finite Layer Neural Network

Neural Network ⟺ Dynamic System

Stochastic Learning ⟺ Stochastic Dynamic System

<u>New Discretization</u>

LM-ResNet

Original One: LM-Resnet56 Beats Resnet110   Modified Equation

Stochastic Depth One: LM-Resnet110 Beats Resnet1202

Thanks For Attention

And Question?

Lu Y, Zhong A, Li Q, et al. Beyond Finite Layer Neural Networks: Bridging Deep Architectures and Numerical Differential Equations arXiv:1710.10121.