DATA DRIVEN METHODS: DEEP RESIDUAL LEARNING, PDE AND BEYOND

YIPING LU 1500010638

Abstract. Due to the success of deep learning, both high level and low level computer vision has caught a lot of people's eye recent years. However there isn't any solid theory to express its success. In this paper, we find out the relationship between PDE and ResNet which are both very deep structure, and have also shown its success in various computer vision works. Based on the scale space theory, there is always a PDE behind vision tasks. We believe that the success of ResNet is based on the its similarity with PDE. We also give some example to utilize PDE to design network, and utilize network to discover new PDEs.

Key Words: Deep Learning, Residual Learning, PDE, Data driven method, Image Processing.

1. Introduction

As living in the era of big data, digital image, no doubt one of most important components of data, have become an important part of our society. With the development of computer and the accumulation of data, many computer vision tasks both low and high level seems can be conquered via learning form huge datasets.

Neural networks has show its power in various computer vision tasks, like denoising, debluring, pattern recognition, object detection, segementation and etc. Among the deep networks, Residual network(ResNet) which was first proposed in He et al.(2015) can achieve lower train and test errors. C Ledig et al.(2017), Zhang et al.(2017) has discorver ResNet's power in image restoration.

Nonlinear PDEs like nonlinear diffusion, is a successful method in low level computer vision tasks like image restoration. Since the introduction of second-order nonlinear diffusion by Perona and Malik in 1990, high order diffusion equations have been proposed widely in low level vision tasks. What these PDE models for image restoration have in common is that they seek a good balance between the two seemingly contradictory objectives: smoothness at locations where noise or other artifacts have been removed, and preservation or even enhancement of the sharpness of edges, corners, etc., which are singularities. Zhao etc al.(2015), Fang etc al.(2017) have also shown that PDE can also been used in pattern recognition and object detection tasks.Moreover The Alvarez-Guichard-Lions-Morel Scale Space Theory shown that there is an PDE behind most of the image processing task. We believe that the success of ResNet may come from its similarity to PDEs and we may utilize ResNet to handle hard PDE tasks.

Instead of learning a function directly, ResNet is proposed to learn a residual in each layer. For forward differential scheme of PDE $u_t = f(u)$, the function $f(u)$ can be consider as the time-wise residual. In order to show this relationship between ResNet and PDE, we developed a new way to design ResNet guided by PDE designing. At the same time, we developed a method to utilize ResNet to discover PDE from data of PDE's solution.

1.1. **Related works.**

1.1.1. *Deep Residual Networks.*

Since first proposed by He et al.(2015)., utilizing identity mapping as the shortcut between different layer of network has become a common trick in network designing. Deep Residual Network as a state-in-art method in deep learning is the deepest structure and brings huge improvements in various computer vision tasks, both low and high level. Recent research shows that ResNet can become more accuracy if we share the weight in different layers and let the network structure wider.

1.1.2. *Data Driven Methods.*

Big data becomes a hot topic recent years, learning from data become a new perspective to design models. The discovery, interpretation and usage of the information, knowledge and resources hidden in all sorts of data to benefit human beings and to improve everyone's day to day life is a challenge to all of us. Deep learning, using deep structure with huge parameters becomes a new and successful method between various machine learning methods. As an example, Cai et al.(2013) developed a data-driven tight frame construction in order to conquer image reconstruction task.

1.1.3. *Wavelet Frame and PDEs.*

Dong et al.(2016) discovers the relationship between the wavelet frames and PDEs, which shows that wavelet may become a better discretization of PDE and PDEs can give geometry background to data-driven wavelet shrinkage.

1.1.4. *Learned Optimal Optimization Via Network Training.*

Back propagation, a way to calculate gradient of complicated functions, makes it possible to learn deep structures. Some iterative methods can be seen as a deep structure with many parameters which need us to give its value at the beginning.Lecun et al.(2010) Xu et al.(2016) using this method to learn the iterative method like ISTA and ADMM to give better performance by the parameter given by human experience. Tompson et al(2016) applied CNN in accelerating Eulerian Fluid Simulation in computer graphics. Many data driven accelerate method has been applied in different problems and has shown its power to perform faster and more accurate.

1.2. **Our Contribution.**

Though Residual Networks and PDEs, both are very deep structure, shown their power in various computer vision tasks, there is not any one to build the bridge between the PDEs and ResNets. First we give a new perspective to understand ResNet via PDEs which is the only way to keep equivariance under rotation and transformation. Next, utilizing the ResNet, we form a new way to discover PDEs form datas, like what the physicians do hundreds of years ago. Different from Liu et al(2010) Hayden(2016), our work doesn't based on a prior dictionary. We also want to show that ResNets can do more than discovering PDEs. The learned filters in ResNet can give advise to numerical methods of PDEs, like intelligent and data-driven finite different scheme designing.

2. Understanding Residual Learning: A PDE Perspective

2.1. **FIR filters and differential operators.**

The concept of vanishing moments and FIR filters is closely related to the orders of differential operators and their corresponding finite difference operators. In this subsection, a key observation is the connection between the vanishing moments of FIR filters and the order of finite difference operators (and the orders of approximation as well).

For an FIR highpass filter $\mathbf{q}$, let $\hat{\mathbf{q}}(\omega) = \sum_{k\in Z^2} \mathbf{q}[k]e^{-ik\omega}$be its two-scale symbol.

Moreover for an FIR highpass filter $\mathbf{q}$, we say it have vanishing moments of order $\alpha = (\alpha_1, \alpha_2)$, where $\alpha \in Z^2_+$ provided that

$$\sum_{k\in Z^2_+} k^\beta \mathbf{q}[k] = i^{|\beta|}\frac{\partial^\beta}{\partial\omega^\beta}\hat{\mathbf{q}}(\omega)|_{\omega=0} = 0$$

for all $\beta \in Z^2$ with $|\beta| < |\alpha|$ and for all $\beta \in Z^2$ with $|\beta| = |\alpha|$ but $\beta \neq \alpha$

Let $\mathbf{q}$ be an FIR highpass filter with vanishing moments of oder $\alpha \in Z^2_+$. Then for a smooth function $F(x)$ on R^2, we have

$$\frac{1}{\delta^{|\alpha|}}\sum_{k\in Z^2} \mathbf{q}[k]F(x+\epsilon k) = C_\alpha \frac{\partial^\alpha}{\partial x^\alpha}F(x) + O(\epsilon)(\epsilon \to 0)$$

Here $C_\alpha = \frac{1}{\alpha!}\sum_k k^\alpha \mathbf{q}[k]$

Proof. The proof is obvious by Taylor Extension. □

2.2. **Understanding ResNet Via PDE.**

. Since the FIR filters may be saw as the discretization of differential operator. The bottle neck structure in ResNet can be seen as

$$u^{(t+1)} = u^t + D_2(S(D_1(u^t)))$$

Here D_2, D_1 are finite schemes of differential operators and the function S can be seen as the active function. Wavelet shrinkage, a famous iterative method in low level computer vision which can be written as $u^{(t+1)} = u^t + W^T(S(W(u^t))$, is the special case of the formula above.Dong et al.(2016) shows that the wavelet shrinkage iterative method converges to a continuous diffusion PDE:$u_t = D^T(f(Du))$. Same as the processing in Dong's work, we can model the ResNet as the PDE below

$$u_t = D_2(S(D_1u))$$

More careful analysis can be seen in the appendix.

2.3. **Nesterov accelerated ResNet.**

As a common trick used to accelerated in optimization, Nesterov accelerated algorithm is widely used in many problems. Su et.al(2015),Andre et.al(2016) shows that the Nesterov accelerated algorithm can be modeled as an ordinary differential equation as

$$\ddot{X}_t + \frac{p+1}{t}\dot{X}_t + Cp^2t^{p-2}\nabla f(X_t) = 0$$

From this observation, if we consider PDE as the gradient flow of a variation problem, a new way to accelerate our PDEs by adding the second order differential term as the formula below is proposed by Osher et al(2010) :

$$\begin{aligned} &\text{Primal Equation:} && u_t = f(u) \\ &\text{Accelerated Equation:} && u_{tt} + cu_t = f(u) \end{aligned}$$

The second order differential term can be consider as the shortcut connection between three layers which is shown in the figure below.

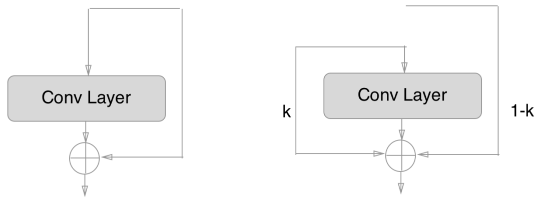

Fig. Newly designed residual block based on Nesterov accelerated PDEs. Left: original residual block used in ResNet, Right:Nesterov accelerated ResNet.

We test our Nesterov ResNet on CIFAR-10 with depth 20 and 32. In every residual block, we follows the design by He et al.(2016)'s which is a 6-layer structure containing two BN layer, two convolutional layer and two ReLu layer. Our model shows better result than the original ResNet with the same depth.

Model	Layer	Parameter	Test Error

Highway	32	1.25M	8.85
RestNet20	20	0.27M	8.75
ResNet 32	32	0.46M	7.51
ResNet 110	32	1.7M	6.43
NResNet 20	20	0.27M	7.53
NResNet 32	32	0.46M	6.93

Table 1: Experiment On Cifar-10,running on one Tesla k4 GPU

3. Discover PDE From Data Via Residual Learning

For PDE is also a type of residual learning, we hope to discover PDE from data via residual learning. The above section has shown that from a learned convolutional kernel, we can calculate its vanishing moment to discover the type of differential operator it belongs to.

3.1. **Multi-Scale Learning.**

To be simple, we consider the linear equation $u_t = \sum_\alpha D^\alpha u$ in this section. Discrete the equation, our finite difference scheme can be written as $u^{(t+1)} - u^t = \sum_\alpha k_\alpha * u^t$ and the task to learn the best scheme can be formula as to learn the best convolutional kernel fits to the data given.

As we all known, the difference scheme can share in different scale. So we apply multi-scale learning to learn the best convolutional kernel that can apply in different scale. To simulate its behavior at lower scale, we apply a stride 2 convolution at the former scale with the same kernel with the formal scale but times $2^{|\alpha|}$ to show the difference of grid size between different scale.

Another reason we apply multi-scale learning is to avoid a problem we called **confusion**. As a simple demo at one dimension case, $[1,-1,0]*u+[0,-1,1]*u = [1,-2,1]*u+[0,0,0]*u$. However we consider the kernel in the LHS as two one-order differential operator but a second order differential operator and a zero operator in the RHS. In order to avoid the phenomena, multi-scale learning is an approach. In order to distinguish different order operators, we should learn the differential equation at different scales.

3.2. **Add Nonlinearity Via RBF.**

. To simplify our model, we ignore the coupled relation between different differential operators. That is to say we only consider the equations like

$$u_t = \sum_{\alpha \in Z_+^2} f_\alpha(D^\alpha u)$$

To discretize the differential operators, we can use a FIR filters(convolutional layer). In order to train the function f_α we use RBF to approximate the function.

Details can be seen in Alexander et al.(2002). In our test, instead of using the tent function, we use the Relu function instead.(Tent function can be seen the sum of some Relu activation after bias translation.)

4. Numerical Experiment

4.1. **Transport Equation.**

We first test a simple example $u_t + u_x = 0$, whose solution is to transport the initial value at time t. Our test shows that the vanish moment of the learned filter is near 0 while higher vanishing moment not, which means our filter is an approximation of a first order differential operator.

Table 2. Learned 3 × 3 Filters From Transport Equation(Without noise)

0.1565	0.2254	-0.3809

Table 3. Learned 5 × 5 Filters From Transport Equation(Without noise)

0.0940	-0.0620	-0.0310	0.4246	-0.4280

4.2. **Heat Equation.**

We also give an example of multi-scale learning method to train the Heat Equation $u_t = \Delta u$. It is well known that the gaussian kernel $\Phi(x,t) = \frac{1}{(4\pi t)^{n/2}} e^{-\frac{|x|^2}{4t}} (t > 0)$ is the solution of Heat Equation with dirac function $\delta(x)$ as the initial boundary condition. We randomly sampled from this solution as our data to discover Heat Equation. In our test, we use three convolution kernel as the zero ,first, second order differential operator and in order to discover them, we trained our data on three scale.Inspired by Schaeffer(2017), we add l1-norm regularization of vanishing moment which means our equation may have simple and sparse formula. In order to test the robustness of our model, in the second test, we add 10% noise. The vanishing moment of our trained filters is shown below. Data below shows that our learned filter have larger vanishing moment at $(2,0),(0,2)$, surprisingly the vanishing moment at $(2,0),(0,2)$ is nearly the same, which means our learned filter can be seen as the approximation of Laplace Operator $\Delta u = u_{xx} + u_{yy}$.

	Vanishing Moment					
Filters	(0,0)	(1,0)	(0,1)	(1,1)	(2,0)	(0,2)
Zero Order	**-0.010496**	-0.004907	0.010003	0.066484	0.014833	0.065231
First Order	-0.006245	**-0.00336**	**0.002333**	0.053965	0.022562	0.033246
Second Order	0.002697	0.019155	0.013385	**-0.020594**	**0.223597**	**0.242161**

Table 4: Learned Filters From Heat Equation(Without Noise)

	Vanishing Moment					
Filters	(0,0)	(1,0)	(0,1)	(1,1)	(2,0)	(0,2)
Zero Order	**-0.012732**	0.007871	0.002985	0.070483	0.061806	0.046284
First Order	0.004880	**0.084576**	**0.066313**	0.276969	0.343246	0.283837
Second Order	-0.000045	0.000144	0.000110	**-0.052367**	**0.174737**	**0.159307**

Table 5: Learned Filters From Heat Equation(With 10% Noise)

4.3. **Diffusion Equation For Image Denoising.**

As an application, we use our framework to train a PDE for Image Denoising. Image denoising is a fundamental operation in image processing and holds considerable practical importance for various real-world applications.We start with conventional nonlinear diffusion processes via ResNet and PDE, then propose a training based reaction diffusion model for image restoration. Perona and Malik diffusion model is a famous model used in image denoiseing which can be shown as

$$\frac{\partial u}{\partial t} = div(c(|\nabla u|^2)\nabla u)$$
$$u(0,x) = u_0(x)$$

Fig.PM equation affter add guassian noise at iterate time:0,15,30,45,60

Via the Isotropic diffusion PDE which is known as one of the smoothing-enhancing PDE, the noise in the image can be smoothed and the edge can be preserved. In our framework, we use an FIR filter convolution layer to approximate the differential operator. It can easily shown that Catte's(1992) approach to solve the Perona-Malik paradox can be shown as

$$\frac{\partial u}{\partial t} = div(c(|\nabla G_\sigma * u|^2)\nabla u)$$
$$u(0,x) = u_0(x)$$

can be easily learned by our learned optimal diffusion model.

To be simple, we ignore the coupled realtion between $\nabla_x u$ and $\nabla_y u$, the P-M model can be also written as

$$\frac{u_t - u_{t-1}}{\Delta t} = \sum_{i=1}^{N_k} K_i^{tT} \phi_i^t (K_i^t u_{t-1}) - \psi(u_{t-1}, f_n)$$

Here we take a reaction term into account, in application we use $\kappa(u_{t-1} - f_n)$ as the reaction term and the parameter κ is also a learnable variable. According to the relation between ResNet and PDEs, every residual block can be plotted as below.

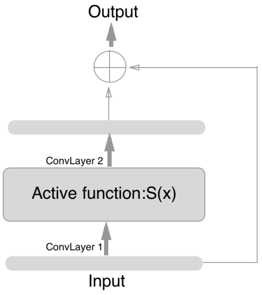

Fig. Every block in learnt optimal diffusion network.

We used the same 400 training images and cropped a 180 × 180 region from each image, resulting in a total of 400 training samples of size 180 × 180, i.e., roughly 13 million pixels. Our train and test data comes from the Berkeley Segmentation Dataset. In our work we test

- 24 filters of size 5 × 5 per layer
- 48 filters of size 7 × 7 per layer
- 24 filters of size 5 × 5 per layer Via multi-scale learning
- 48 filters of size 7 × 7 per layer Via multi-scale learning

We trained our model via Adam and Sgd utilizing weight decay with batch size 10 on a Tesla K40 GPU, the testing result and comparison with the state-of-art methods is shown below.

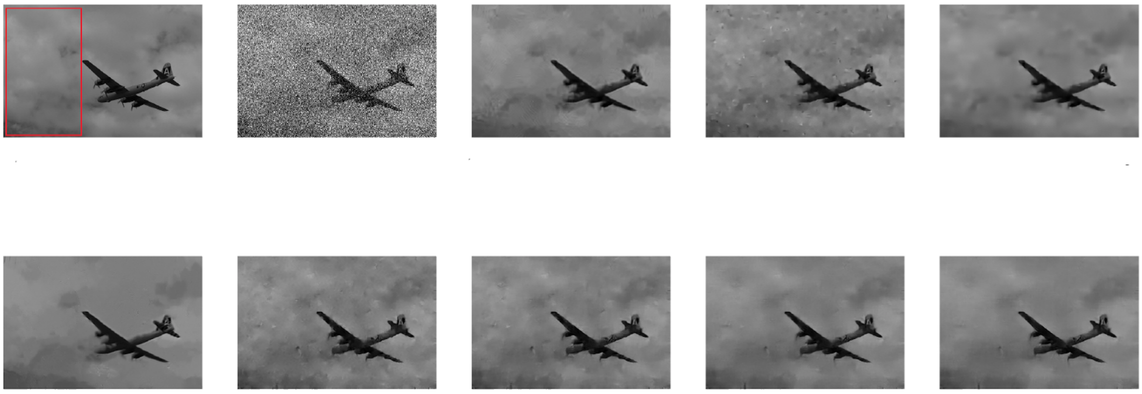

Fig. Gaussian Noise Denoising Test:(a) clean image(b) noised image (c) BM3D (d) EPLL (e) SRCNN (f) WNNM (g)5×5 filters (h) 7×7 filters (i) 5×5 filters with multi-scale learning (j)7×7 filters with multi-scale learning

5. Further Work.

According to the relationship between the PDEs and ResNets can give new insight in both PDE and deep learning.

5.1. **Application in low level vision tasks.** There are more tasks in low level vision tasks can be solved by PDEs like image inpainting, image blind debluring and etc. Traditional methods may not work sometimes. But what if we using our training methods? The success in image denoising gives us confidence to conquer more tasks in low level vision tasks.

5.2. **Discover PDEs in high level vision tasks.** Designing PDEs for image vision tasks needs strong geometry insights of the problems, but sometimes it is hard to give geometry explanation to some complicated tasks like face recognition. In our learning method, PDEs can be learned from the huge amount of datas which is easy to available.

5.3. **Conquer more physical problems and PDEs.** Discovering PDEs from data is a traditional method in physical. By the development of computer, this work may be done automatically, which can help us to have deep insight to physical phenomena. At the same time there is also a lot PDEs hard to solve, we hope to discover a method to solve the PDEs by learning from the data.Our learned FIR filters can be new way finite differential scheme designed data driven. This may also gives deep insight into numerical PDEs.

5.4. **Give advise to Network structure.** We have already many results in PDEs, but we still have less interpretation in network designing. Like Nesterov ResNet and optimal diffusion network we design, we believe we can find network structure via PDE. At the same time, using PDEs can add more prior knowledge into network designing which can become more accurate and need less data at the same time.

6. Conclusion

In our paper, we consider ResNet as a method to utilize a network structure to learn a PDE $u_t = f(u)$. In this perspective, we may understand the reason ResNet can become deeper and deeper. From PDE perspective, ResNet can have many changes, we give an example to show that Nesterov accelerate can be applied in network structure designing. At the same time, ResNet can bring new perspective to develop data driven methods in numerical PDE.

7. Acknowledgement

I would like to thank Xianzong Ma and Zichao Long for many useful discussion. I also thank Prof.Bin Dong for recently drawing my attention this project, giving me a lot of help and instruction and providing GPU for our test. Finally, my thank would also gives to Prof. Pingwen Zhang, who taught me the way to do the research during the course 'learning by researching'.

8. Reference

He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// Computer Vision and Pattern Recognition. IEEE, 2015:770-778.

Chen Y, Yu W, Pock T. On learning optimized reaction diffusion processes for effective image restoration[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2015:87-90.

Bin Dong, Qingtang Jiang and Zuowei Shen, Image restoration: wavelet frame shrinkage, nonlinear evolution PDEs, and beyond, Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal, 15(1)?606-660, 2017.

Bin Dong, Zuowei Shen and Peichu Xie, Image restoration: a general wavelet frame based model and its asymptotic analysis, SIAM Journal on Mathematical Analysis, 49(1), 421-445, 2017.

Jian-Feng Cai, Bin Dong and Zuowei Shen, Image restoration: a wavelet frame based model for piecewise smooth functions and beyond, Applied and Computational Harmonic Analysis, 41(1), 94-138, 2016.

Jian-Feng Cai, Bin Dong, Stanley Osher and Zuowei Shen, Image restoration: total variation; wavelet frames; and beyond, Journal of the American Mathematical Society, 25(4), 1033-1089, 2012.

Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553):436-444.

Wibisono A, Wilson A C, Jordan M I. A variational perspective on accelerated methods in optimization.[J]. Proceedings of the National Academy of Sciences of the United States of America, 2016, 113(47):E7351.

Su W, Boyd S, Cands E J. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights[C]// International Conference on Neural Information Processing Systems. MIT Press, 2014:2510-2518.

Fang C, Zhao Z, Zhou P, et al. Feature Learning via Partial Differential Equation with Applications to Face Recognition[J]. Pattern Recognition, 2017.

Zhang K, Zuo W, Chen Y, et al. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising[J]. IEEE Transactions on Image Processing, 2017, PP(99):1-1.

Ledig C, Theis L, Huszar F, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network[J]. 2017.

Feng W, Qiao P, Xi X, et al. Image Denoising via Multi-scale Nonlinear Diffusion Models[J]. 2016.

Gregor K, Lecun Y. Learning Fast Approximations of Sparse Coding[C]// Proc. International Conference on Machine Learning. 2010.

Y Yang, J Sun, H Li, Z Xu Deep ADMM-Net for Compressive Sensing MRI In Advances in Neural Information Processing Systems 2016.

Jin K H, Mccann M T, Froustey E, et al. Deep Convolutional Neural Network for Inverse Problems in Imaging[J]. 2016.

Xin B, Wang Y, Gao W, et al. Maximal Sparsity with Deep Networks?[J]. 2016.

Liu R, Lin Z, Zhang W, et al. Learning PDEs for image restoration via optimal control[C]// European Conference on Computer Vision. Springer-Verlag, 2010:115-128.

H Schaeffer Learning partial differential equations via data discovery and sparse optimization. 2017

Bin Dong and Zuowei Shen, Image restoration: a data-driven perspective, Proceedings of the International Congress of Industrial and Applied Mathematics (ICIAM), Beijing, China, High Education Press (Lei Guo and Zhi-Ming Ma eds), 65-108, 2015.

U. Schmidt and S. Roth. Shrinkage fields for effective image restoration. In CVPR, 2014

Kichenassamy S. The Perona-Malik Paradox[J]. Siam Journal on Applied Mathematics, 1997, 57(5):1328-1342.

Francine, Lions P L, Morel J M, et al. Image selective smoothing and edge detection by nonlinear diffusion[J]. Siam Journal on Numerical Analysis, 1992, 29(1):182-193.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In ICCV, pages 2272?2279, 2009

Osher S, Yu M, Dong B, et al. Fast Linearized Bregman Iteration for Compressive Sensing and Sparse Denoising[J]. Mathematics of Computation, 2011, 8(1):93-111.

Cai J F, Ji H, Shen Z, et al. Data-driven tight frame construction and image denoising[J]. Applied & Computational Harmonic Analysis, 2013, 37(1):89-105.

K. Niklas Nordstrom. Biased anisotropic diffusion: a unified regular- ? ization and diffusion approach to edge detection. Image and Vision Computing, 8(4):318?327, 1990

P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. IEEE TPAMI, 12(7):629?639, 1990

Rigelsford J. Handbook of Neural Network Signal Processing[J]. Industrial Robot, 2003, 111(4):780-780.

Tompson J, Schlachter K, Sprechmann P, et al. Accelerating Eulerian Fluid Simulation With Convolutional Networks[J]. 2016.

Evans L C. Partial differential equations[M]. AMS,Graduate Studies in Mathematics 2009.

V. T. Finite Difference Schemes and Partial Differential Equations. by John C. Strikwerda[M]// Finite difference schemes and partial differential equations. Wadsworth & Brooks Cole Advanced Books & Software, 1989:xii,435.

Daubechies I, Heil C. Ten Lectures on Wavelets[J]. Journal of the Acoustical Society of America, 1992, 93(3):1671.

Stephane. A Wavelet Tour of Signal Processing[M]. Academic press, 1998.

Bin Dong and Zuowei Shen, MRA-based wavelet frames and applications, IAS Lecture Notes Series, 2010 Summer Program on "The Mathematics of Image Processing", Park City Mathematics Institute, expected to appear in 2012 (manuscript revised on Sep 28, 2011).

Aubert G, Kornprobst P. Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations (Applied Mathematical Sciences)[J]. Applied Intelligence, 2006, 40(2):291-304.

Appendix. Mathematical Analysis Of Relationship Between PDEs And Networks

In this section we consider nonlinear evolution PDEs $u_t = \sum_{l=1}^{L} \frac{\partial^{\alpha_l}}{\partial x^{\alpha_l}} \Phi_l(Du, u)$, with $D = (\frac{\partial^{\beta_1}}{\partial x^{\beta_1}}, \cdots, \frac{\partial^{\beta_l}}{\partial x^{\beta_l}})$. Many PDE in image processing can seen as the example of nonlinear evolution PDEs:

- PM equation: $u_t = div(g(|\nabla u|^2)\nabla u)$
- Osher and Rudin Shock filter: $u_t + |\nabla u| \, F(L(u)) = 0$

Now we give a bi-frame filter bank $\left\{q^{(1)}, q^{(2)}, \cdots, q^{(L)}\right\}$ and $\left\{\widetilde{q^{(1)}}, \widetilde{q^{(2)}}, \cdots, \widetilde{q^{(L)}}\right\}$, A one residual block in ResNet can be formulated as

$$u_j^i = u_j^0 + \sum_{l=1}^{L} \widetilde{q^{(i)}}\,[j-n]\,(\Phi_i(\xi))|_{\xi = H_n^{(l)}}$$

Here $H_n^{(l)} = \sum_{j \in Z^2} q^{(l)}[j] u_{j+n}^0$

Here we give another definition, we say a FIR filter $\mathbf{q}$ have total vanishing moment $K \setminus \{J+1\}$ if

$$\sum_{k \in Z_+^2} k^{\beta} \mathbf{q}[k] = i^{|\beta|} \frac{\partial^{\beta}}{\partial \omega^{\beta}} \hat{\mathbf{q}}(\omega)|_{\omega=0} = 0$$

for all $\beta \in Z_+^2$ with $|\beta| < K$ except for $\beta \neq \beta_0$ with certain $\beta_0 \in Z_+^2$ and $|\beta_0| = J < K$

With the notation of total vanishing moment, it is easy to estimate the local truncation error as $O(K-J)$, which has been proved by Dong(2016).

Difference from Finite difference approximation For finite difference approximation method can be seen as $\mathbf{u} = R_h u$, using filter ψ is another way to do the sample. From the observation $\left\langle u, D^T \phi_{n,k} \right\rangle = \langle u, \psi_{n,k} \rangle$, we use $\psi_{n,k}$ to sample u instead of point-wise sample, which only need our signal to be week differentiable but not smooth. At the same time, some filters like WFT(wavelet frame transform) may have a left inverse. Compare with wavelet Galerkin method $\left\langle u, D^T \phi_{n.k} \right\rangle$, we doesn't need to calculate the connection coefficients, which is not as convenient and simple as using WFT and filter convolution.