Geometric Understanding of Supervised and Unsupervised Deep Learning for Biomedical Image Reconstruction

Jong Chul Ye, Ph.D

Professor

\textit{BISPL - BioImaging, Signal Processing, and Learning Lab.}

KAIST, Korea
Classical Learning vs Deep Learning

Classical machine learning

Feature Engineering

Deep learning (no feature engineering)

Deep Learning Era in Medical Imaging

Diabetic eye diagnosis
Gulshan, V. et al. JAMA (2016)

Skin Cancer diagnosis

Image segmentation
Ronneberger et al, MICCAI, 2015

OCT diagnosis

Chest X-ray
Courtesy of Kyu Hwan Jung @Vuno

Image registration
Figure courtesy of X. Cao & D. Shen
Deep Learning for Image Reconstruction

Focus of this talk: Image reconstruction

Diagnosis & analysis
Deep Learning Revolution for Inverse Problem

- **Accuracy**: high quality recon > CS
- **Fast** reconstruction time
- **Business model**: vendor-driven training
- **Interpretable** models

<table>
<thead>
<tr>
<th></th>
<th>Imaging time</th>
<th>Reconstruction time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compressed Sensing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine Learning</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DOES IT CREATE ANY ARTIFICIAL FEATURES?
Geometry of Supervised Learning

Ye et al, SIAM J. Imaging Sciences, 2018; Ye et al, ICML, 2019
Encoder-Decoder CNN for Inverse Problems
Encoder-Decoder CNN for Inverse Problems
Successful applications to various inverse problems
Why *Same* Architecture Works for *Different* Inverse Problems?
Step 1: Signal Representation

\[ x = \sum_i \langle b_i, x \rangle \tilde{b}_i \]
Classical Methods for Inverse Problems

Step 2: Basis Pursuit by Optimization

Eg. Compressed Sensing

\[
x = \sum_i \tilde{b}_i \langle b_i, x \rangle
\]
Why do They Look so Different?
Any Link between Them?
Our Theoretical Findings

Ye et al, SIIMS, 2018; Ye et al, ICML, 2019

\[ y = \sum_i \langle b_i(x), x \rangle \tilde{b}_i(x) \]
Our Theoretical Findings

Ye et al, SIIMS, 2018; Ye et al, ICML, 2019

\[ y = \sum_i \langle b_i(x), \tilde{b}_i(x) \rangle \]
Our Theoretical Findings

Ye et al, SIIMS, 2018; Ye et al, ICML, 2019

\[
y = \sum_i \langle b_i(x), x \rangle b_i(x)
\]
Our Theoretical Findings

Ye et al, SIIMS, 2018; Ye et al, ICML, 2019

\[ y = \sum_i \langle b_i(x), x \rangle b_i(x) \]
Our Theoretical Findings

Ye et al, SIIMS, 2018; Ye et al, ICML, 2019

\[
y = \sum_i \langle b_i(x), x \rangle \tilde{b}_i(x)
\]
Linear Encoder-Decoder (ED) CNN

\[ y = \tilde{B} B^\top x = \sum_i \langle x, b_i \rangle \tilde{b}_i \]

\[
B = E^1 E^2 \cdots E^\kappa,
\]
\[
\tilde{B} = D^1 D^2 \cdots D^\kappa
\]

**Pooling**

\[ E^l = \begin{bmatrix}
\tilde{\Phi}^l \otimes \psi^l_{1,1} & \cdots & \tilde{\Phi}^l \otimes \psi^l_{1,q_l,1} \\
\vdots & \ddots & \vdots \\
\tilde{\Phi}^l \otimes \psi^l_{1,q_l,q_l-1} & \cdots & \tilde{\Phi}^l \otimes \psi^l_{1,q_l,q_l}
\end{bmatrix} \]

**Un-pooling**

\[ D^l = \begin{bmatrix}
\tilde{\Phi}^l \otimes \tilde{\psi}^l_{1,1} & \cdots & \tilde{\Phi}^l \otimes \tilde{\psi}^l_{1,q_l,1} \\
\vdots & \ddots & \vdots \\
\tilde{\Phi}^l \otimes \tilde{\psi}^l_{q_l-1,1} & \cdots & \tilde{\Phi}^l \otimes \tilde{\psi}^l_{q_l-1,q_l}
\end{bmatrix} \]

Learned filters
Linear E-D CNN w/ Skipped Connection

\[ y = \tilde{B}B^\top x = \sum_i \langle x, b_i \rangle \tilde{b}_i \]

\[
B = \begin{bmatrix} E^1 & \cdots & E^\kappa \\ \end{bmatrix} \quad \tilde{B} = \begin{bmatrix} D^1 & \cdots & D^\kappa \\ \end{bmatrix}
\]

\[
E^1 \cdots E^{\kappa-1} S^\kappa \quad \cdots \quad E^1 S^2 \quad S^1
\]

\[
D^1 \cdots D^{\kappa-1} \tilde{S}^\kappa \quad \cdots \quad D^1 \tilde{S}^2 \quad \tilde{S}^1
\]

more redundant expression

\[
S^l = \begin{bmatrix} I_{m_{l-1}} \otimes \psi_{1,1}^l & \cdots & I_{m_{l-1}} \otimes \psi_{q_{l-1},1}^l \\ \vdots & \ddots & \vdots \\ I_{m_{l-1}} \otimes \psi_{1,q_{l-1}}^l & \cdots & I_{m_{l-1}} \otimes \psi_{q_{l-1},q_{l-1}}^l \\ \end{bmatrix}
\]

\[
\tilde{S}^l = \begin{bmatrix} I_{m_{l-1}} \otimes \tilde{\psi}_{1,1}^l & \cdots & I_{m_{l-1}} \otimes \tilde{\psi}_{q_{l-1},1}^l \\ \vdots & \ddots & \vdots \\ I_{m_{l-1}} \otimes \tilde{\psi}_{1,q_{l-1}}^l & \cdots & I_{m_{l-1}} \otimes \tilde{\psi}_{q_{l-1},q_{l-1}}^l \\ \end{bmatrix}
\]

Learned filters
Deep Convolutional Framelets

Perfect reconstruction

\[ x = \tilde{B} B^\top x = \sum_i \langle x, b_i \rangle \tilde{b}_i \]

Frame conditions

\( \Psi^l \tilde{\Psi}^l = \frac{1}{r\alpha} I_{r q_i - 1} \)

\( \Phi^l \tilde{\Phi}^l = \alpha I_{m_{l-1}} \)

w/o skipped connection

w skippied connection
Deep Convolutional Framelets

Perfect reconstruction

\[ x = \tilde{B} B^\top x = \sum_i \langle x, b_i \rangle \tilde{b}_i \]

Frame conditions

\begin{align*}
\text{w/o skipped connection} & \quad \tilde{\Phi}^l \Phi^l \top = \alpha I_{m_{l-1}} \quad \text{and} \quad \Psi^l \tilde{\Psi}^l \top = \frac{1}{r \alpha} I_{r q_{l-1}} \\
\text{w skipped connection} & \quad \tilde{\Phi}^l \Phi^l \top = \alpha I_{m_{l-1}} \quad \text{and} \quad \Psi^l \tilde{\Psi}^l \top = \frac{1}{r (\alpha + 1)} I_{r q_{l-1}}
\end{align*}

Ye et al, SIAM J. Imaging Science, 2018
Role of ReLUs?
Generator for Multiple Expressions

\[ y = \tilde{B}(x)B(x)^\top x = \sum_i \langle x, b_i(x) \rangle \tilde{b}_i(x) \]

\[
\begin{align*}
B(x) &= E^1 \Sigma^1(x) E^2 \cdots \Sigma^{\kappa-1}(x) E^\kappa, \\
\tilde{B}(x) &= D^1 \Sigma^1(x) D^2 \cdots \Sigma^{\kappa-1}(x) D^\kappa
\end{align*}
\]

\[
\Sigma^i(x) = \begin{bmatrix}
\sigma_1 & 0 & \cdots & 0 \\
0 & \sigma_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma_{m_i}
\end{bmatrix}
\]

Input dependent \{0,1\} matrix

--> Input adaptivity
Input Space Partitioning for Multiple Expressions
A CNN performs automatic assignment of distinct linear representation depending on input.
Expressivity of E-D CNN

# of representation

# of network elements
Expressivity of E-D CNN

# of network elements vs. # of representation

# of channel
Expressivity of E-D CNN

# of representation

# of network elements

# of channel

Network depth
Expressivity of E-D CNN

Skipped connection

# of channel

Network depth

# of representation

# of network elements
Lipschitz Continuity

Related to the generalizability

\[ \| F(W, x^{(1)}) - F(W, x^{(2)}) \|_2 \leq K \| x^{(1)} - x^{(2)} \|_2 \]

\[ K = \max_p K_p, \quad K_p = \| \tilde{B}(z_p) B(z_p)^\top \|_2 \]

Dependent on the Local Lipschitz
Regularized Recon vs. Deep Recon

Classical Regularized Recon (basis engineering)

Deep Recon (no basis engineering)
THEORY-DRIVEN CNN DESIGN
:some snapshots
Universal Deep Beamformer

Khan et al, MICCAI, 2019 (oral presentation)
Ultrasound Acquisition Modes

Planewave Imaging

Focused Imaging

(a) Beam Emission

(b) RFs Record

(c) Beam Sweeping
Conventional approach

Conventional *delay-and-sum* (DAS) Beamforming Pipeline

- **Rx** probe
- **Beam focusing**
- **Signal Adder**
- **Hilbert Transform**
- **Envelope detection**
- **Log compression**

**IQ data**
*(signal before envelope detection)*

**B-mode image**
Adaptive Beamformer

Conventional Beamforming Pipeline

\[
\begin{bmatrix}
I \\
Q
\end{bmatrix} = \begin{bmatrix}
z_l[n] \\
\mathcal{H}(z_l)[n]
\end{bmatrix}
\]

Data specific weights

\[
z_l[n] = w_l[n]^\top y_l[n] \quad w_l[n] = \frac{R_l[n]^{-1}1}{1^H R_l[n]^{-1}1}
\]
Adaptive and Compressive Deep Beamformer

Conventional Beamforming Pipeline

- Input partitioning
  → Data specific representation

Deep Beamformer

Khan et al, MICCAI 2019
Universal Deep Beamformer

Convolutional neural network (CNN)

IQ data
*(signal before envelope detection)*

B-mode image
Application to Planar Wave Imaging

Data Generation

Proposed Method

Time of flight

Planewave1

Planewave2

Planewave3

Final image

Rx

Plane wave Rx data cube
Results (Planewave / in-vivo)

<Full data>
31 plane waves and 64 channels

3 plane waves and 64 channels

31 plane waves and 32 channels

31 plane waves and 8 channels

SSIM: 1
CNR: 2.81

SSIM: 0.70
CNR: 2.23

SSIM: 0.86
CNR: 2.31

SSIM: 0.66
CNR: 1.73

SSIM: 0.99
CNR: 2.83

SSIM: 0.73
CNR: 2.57

SSIM: 0.89
CNR: 2.88

SSIM: 0.77
CNR: 2.69

Lateral length (mm)

Depth (mm)

DAS

DeepBF

Invivo CNR (PW Sampling)

Invivo CNR (Random Sampling)
Application to Focused B-Mode Imaging

Data Generation

Focusing

Scanline

Depth

Rx

Proposed Method

Focused Rx data cube

Input

Output

IQ Data

module1 module2 module3 module4 module5 module6 module7 module8 module9

3x3 kernel convolution, batch norm, ReLU

1x1 kernel convolution

Sub-cube selection

contracting path, concatenate
Results of Deconvolution BF (phantom)
Results of Deconvolution BF (in-vivo)
Which domain is good for learning?

Han et al, IEEE Trans. Medical Imaging (in press), 2019  
Lee et al, MRM (in press), 2019  
Han et al, Medical Physics, 2020
Image Domain Learning is Essential?

Kravitz et al, Trends in Cognitive Sciences January 2013, Vol. 17, No. 1
k-Space Deep Learning

Han et al, IEEE Trans. Medical Imaging (in press), 2019
Lee et al, MRM (in press), 2019
K-space Deep Learning (Radial R=6)

Ground-truth

Image learning

Acceleration

CS

K-space learning

Han et al, *IEEE TMI* (in press)
K-space Deep Learning (Radial R=6)

Ground-truth

Image learning

Acceleration

CS

K-space learning

Han et al, *IEEE TMI* (in press)
DBP Domain Deep Learning

Differentiated Backprojection

\[ g(r') = - \frac{1}{2\pi} \int_{\lambda_1}^{\lambda_2} \frac{d\lambda}{|r' - r_0(\lambda)|} \left[ \frac{\partial}{\partial q} p(r_0(q), \beta(r', \lambda)) \right]_{q=\lambda} \]

Han et al, Medical Physics 46 (12), e855-e872, 2020
Two Approaches for CT Reconstruction


Filtered Backprojection (FBP)

- Ramp filtering
- Back-projection

Backprojection Filtration (BPF)

- Differentiation
- Back-projection
- Hilbert transform

\[
f(u) = \frac{c}{\pi \sqrt{1 - u^2}} - \frac{1}{\pi \sqrt{1 - u^2}} \int_{-1}^{1} \frac{\sqrt{1 - s^2} g(s)}{u - s} ds,
\]
DBP Domain ROI Tomography

Interior (ROI) Tomography → 2-D Deconvolution problem

Han et al, Medical Physics, 2019
DBP Domain Conebeam Artifact Removal

Han et al, arXiv:1906.06854

Standard Method: FDK Algorithm

https://www.ndt.net/article/wcndt00/papers/idn730/idn730.htm
DBP Domain Conebeam Artifact Removal

Han et al, arXiv:1906.06854

Exact Factorization $\rightarrow$ 2-D Deconvolution problem

$$g(t, z) = \pi \int_{-\infty}^{\infty} h_H(t - \tau) (f(\tau, z_1(\tau)) + f(\tau, z_2(\tau))) d\tau$$
Improving U-Net

Ye et al, SIAM J. Imaging Science, 2018
Han et al, IEEE Trans. Medical Imaging, 2018
Yoo et al, SIAM J. Applied Math, 2019
Limitation of U-Net

\[ \Phi = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \\ 0 & 0 & \ldots & 1 \end{bmatrix} \]

U-Net Pooling does NOT satisfy the frame condition

\[ \Phi_{\text{ext}} \Phi_{\text{ext}}^T = I + \Phi \Phi^T \neq I \]

JC Ye et al, SIAM Journal Imaging Sciences, 2018
Y. Han et al, TMI, 2018.
Improving U-net by Frame Conditions

- **Dual Frame U-net**

- **Tight Frame U-net**

JC Ye et al, SIAM Journal Imaging Sciences, 2018
Y. Han and J. C. Ye, TMI, 2018
U-Net vs Tight-Frame U-Net

Y. Han and J. C. Ye, TMI, 2018; Yoo et al, SIJAM, 2018
Photorealistic Style Transfer via Wavelet Transforms

Jaejun Yoo* Youngjung Uh* Sanghyuk Chun* Byeongkyu Kang Jung-Woo Ha
Clova AI Research, NAVER Corp.

Multi-level stylization

(a) WCT (artistic)  
(b) PhotoWCT  
(c) Ours (WCT²)

Progressive stylization

WCT  
Max-pooling / Upsampling  
Unpooling  
Wavelet pooling  
Wavelet unpooling

ICCV 2019
What if we do not have label data for training?
Yann LeCun’s Cake Analogy

- **“Pure” Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**

- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input.
  - Predicting human-supplied data
  - **10→10,000 bits per sample**

- **Unsupervised/Predictive Learning (cake)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

(Yes, I know, this picture is slightly offensive to RL folks. But I’ll make it up)
Motivation

• **Multiphase Cardiac CT denoising**
  - Phase 1, 2: low-dose, Phase 3 ~ 10: normal dose
  - Goal: dynamic changes of heart structure
  - No reference available

Kang et al, Medical Physics, 2018
Unsupervised Denoising for Low-Dose CT

Kang et al, Medical Physics, 2019
Lose dose (5%) $\rightarrow$ high dose
Ablation Study

Input: phase 1

Proposed

Without identity loss

GAN

Kang et al, Medical Physics, 2018
Ablation Study

Input: phase 1  Proposed  Without identity loss  GAN

(a)  (b)  (c)  (d)

(e)  (f)  (g)  (h)

Kang et al, Medical Physics, 2018
Geometry of CycleGAN

Sim et al, arXiv:1909.12116, 2019
Lim et al, arXiv:1908.09414, 2019
Kim et al, MICCAI, 2019
Kang et al, Medical Physics, 2019
Optimal Transport (OT)

Kantorovich Formulation: minimizing average transport cost

$$\min_{\pi} \int_{\mathcal{X} \times \mathcal{Y}} \|x - y\| d\pi(x, y)$$

Transport cost

Villani, 2008; Peyre et al, 2019
Optimal Transport (OT)
Villani, 2008; Peyre et al, 2019

Wasserstein GAN: if transport cost is given by \( c(x, y) = \|x - y\| \)

\[
\min_{\Theta} \max_{\varphi} \int_{\mathcal{X}} \varphi(x) d\mu(x) - \int_{\mathcal{Y}} \varphi(G_{\Theta}(y)) d\nu(y)
\]

1-Lipschitz function
Our Theoretical Findings
Sim et al, arXiv:1909.12116, 2019

Optimal transport: Kantorovich Formulation

$$\min_{\pi} \int_{X \times Y} c(x, y) \, d\pi(x, y)$$
Our Penalized LS Formulation

CNN regularization

\[ c(x; y, \Theta, \mathcal{H}) = \| y - \mathcal{H}x \|^q + \| G_\Theta(y) - x \|^p \]
Our Theoretical Findings
Sim et al, arXiv:1909.12116, 2019

Optimal transport with PLS as transport cost

$$\min_{\pi} \int_{\mathcal{X} \times \mathcal{Y}} \|y - \mathcal{H}x\|^q + \|G_{\Theta}(y) - x\|^p \, d\pi(x, y)$$

Both $x, y$ are stochastic: stochastic generalization of PLS
Our Theoretical Findings
Sim et al, arXiv:1909.12116, 2019

Under some regularity conditions, Kantorovich Dual Formulation $\rightarrow$ CycleGAN

\[
\min_{\Theta, \Phi} \max_{h, \Xi} \ell_{\text{cycle}}(\Theta, \mathcal{H}) + \ell_{\text{GAN}}(\Theta, \mathcal{H}; \Phi, \Xi)
\]

\[
\ell_{\text{cycle}}(\Theta, \mathcal{H}) = \rho \int_{\mathcal{X}} \|x - G_{\Theta}(\mathcal{H}x)\|^p d\mu(x) + \sigma \int_{\mathcal{Y}} \|y - \mathcal{H}G_{\Theta}(y)\|^q d\nu(y)
\]

\[
\ell_{\text{GAN}}(\Theta, \mathcal{H}; \Phi, \Xi)
\]

\[
= \int_{\mathcal{X}} \varphi_{\Phi}(x) d\mu(x) - \int_{\mathcal{Y}} \varphi_{\Phi}(G_{\Theta}(y)) d\nu(y) + \int_{\mathcal{Y}} \psi_{\Xi}(y) d\nu(y) - \int_{\mathcal{X}} \psi_{\Xi}(\mathcal{H}x) d\nu(x)
\]
THEORY-DRIVEN CYCLEGAN DESIGN : some snapshots
Unsupervised Blind Deconvolution Microscopy

\[ c(x, y; \Theta, h) = \| y - h \ast x \| + \| G_{\Theta}(y) - x \| \]
Results on Real Microscopy Data

✓ Qualitative results

- Runtime for 512 x 512 x 50 volume inference: 15 s

Lim et al, arXiv:1908.09414, 2019
Unsupervised Learning for Accelerated MRI

\[ c(x, y; \Theta) = \| y - F^{-1} P_{\Omega} F x \| + \| G_{\Theta}(y) - x \| \]
Results on Fast MR Data Set

Sim et al, arXiv:1909.12116, 2019
Summary

• Deep learning has become an important platform for medical imaging

• Our theoretical findings
  • Deep neural network with ReLU is a piecewise linear frame representation
  • Unsupervised learning can be solved with optimal transport with PLS cost == cycleGAN

• We can design problem-specific neural networks
Acknowledgement

- Daniel Rueckert (Imperial College)
- Florian Knoll (NYU)
- Fang Liu (Univ. of Wisconsin)
- Mehmet Akcakaya (Univ. of Minnesota)
- Dong Liang (SIAT, China)
- Dinggang Shen (UNC)
- Peder Larson (UCSF)
- Hyunwook Park (KAIST)
- Sung-hong Park (KAIST)
- Jongho Lee (SNU)
- Doshik Hwang (Yonsei Univ)
- Won-Jin Moon (KonkukUniv Medical Center)
- Eungyeop Kim (Gachon Univ. Medical Center)
- Leonard Sunwoo (SNUBH)
- Kyuhwan Jung (Vuno)

- Grant
  - NRF of Korea
  - Ministry of Trade Industry and Energy