EPFL

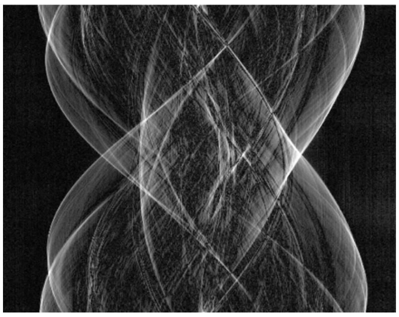
CIBM
Center for Biomedical Imaging

Section: Mathematical Imaging

Part III: Deep splines and the Robust Learning of Nonlinearities

Michael Unser

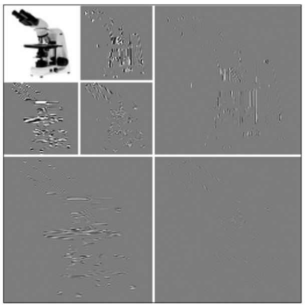
Biomedical Imaging Group
& EPFL Center for Imaging



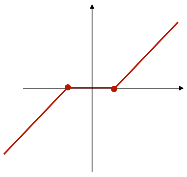
Summer School, Mathematics and Machine Learning for Image Analysis, Bologna, June 4-12, 2024

The building blocks of classical image processing

- Linear transforms
  - Digital filters
  - Fourier transform (FFT)
  - Wavelet transform, DCT
  - Karhunen-Loève tranform
  - Independent component analysis



- Pointwise non-linearities
  - Gain control
  - Thresholding, clipping
  - Soft-threshold



The building blocks of classical and modern image processing

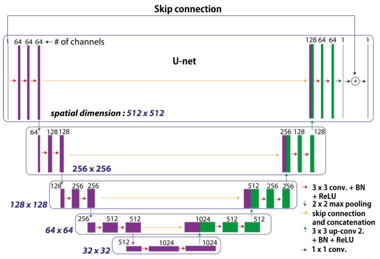
- Linear transforms
  - Digital filters
  - Fourier transform (FFT)
  - Wavelet transform
  - Karhunen-Loève tranform
  - Independent component analysis
- Pointwise non-linearities
  - Gain control
  - Thresholding, clipping
  - Soft-threshold
- Specialized hardward: GPUs
- Integrated software frameworks

PyTorch TensorFlow

Neural networks

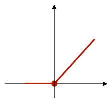
Linear weights

- Fully connect layers
- Convolutional layers
- Multi-channel filterbanks

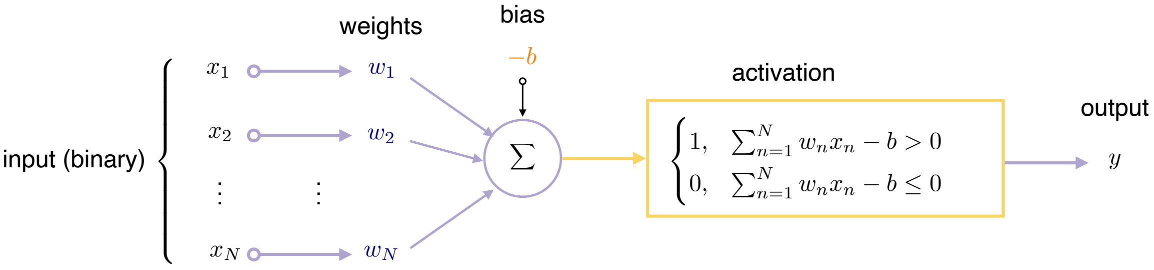


Activation functions

- Sigmoid
- ReLU



Formal model of neuron (McCulloch & Pitt)



weights

bias

$-b$

activation

output

input (binary)

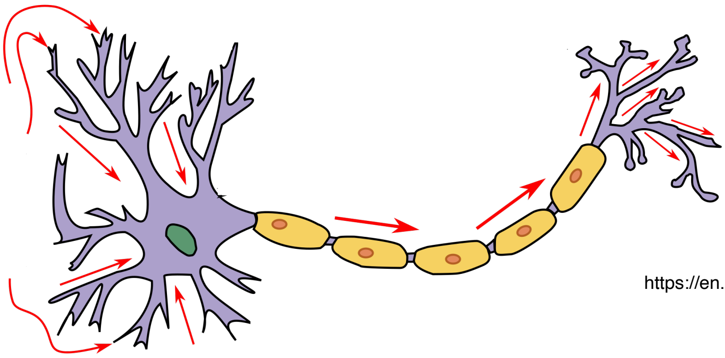
x_1, x_2, $\vdots$, x_N

w_1, w_2, $\vdots$, w_N

Σ

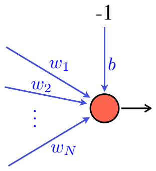
$$\begin{cases} 1, & \sum_{n=1}^{N} w_n x_n - b > 0 \\ 0, & \sum_{n=1}^{N} w_n x_n - b \leq 0 \end{cases}$$

y



https://en.wikipedia.org/wiki/Artificial_neuron

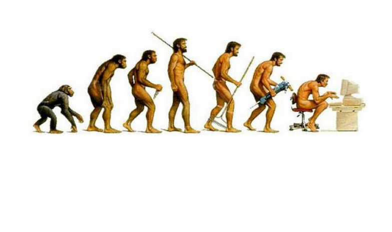
Artificial neurons



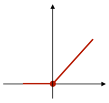
Definition: An artificial neuron with weights $\mathbf{w} = (w_1, \dots, w_N) \in \mathbb{R}^N$, bias $b \in \mathbb{R}$ and **activation function** $\sigma : \mathbb{R} \to \mathbb{R}$ is defined as the function $f : \mathbb{R}^N \to \mathbb{R}$

$$f(\boldsymbol{x}) = \sigma\left(\mathbf{w}^\mathsf{T}\boldsymbol{x} - b\right) = \sigma\left(\sum_{n=1}^{N} w_n x_n - b\right).$$

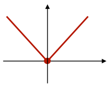
- Examples of activation functions
  - Threshold Logic Unit (Heaviside): $\mathrm{TLU}(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$ (McCullogh & Pitt 1943; Rosenblatt 1957)
  - Sigmoid function: $\sigma(x) = \dfrac{1}{1 + \mathrm{e}^{-x}}$ (Rumelhart 1986, ...)

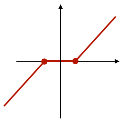


  - Rectified Linear Unit: $\mathrm{ReLU}(x) = x_+ = \max(0, x)$



  - And variants .....





The building blocks of classical and modern image processing

- Linear transforms
  - Digital filters
  - Fourier transform (FFT)
  - Wavelet transform
  - Karhunen-Loève tranform
  - Independent component analysis
- Pointwise non-linearities
  - Gain control
  - Thresholding, clipping
  - Soft-threshold
- Integrated software framework

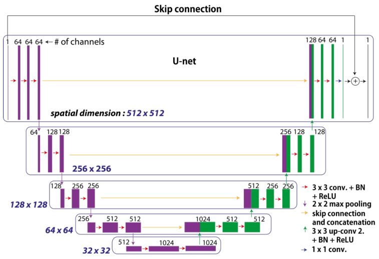
Neural networks

Linear weights

- Fully connect layers
- Convolutional layers
- Multi-channel filterbanks

Trainable

Activation functions



Why & how

NEW

PyTorch

OUTLINE

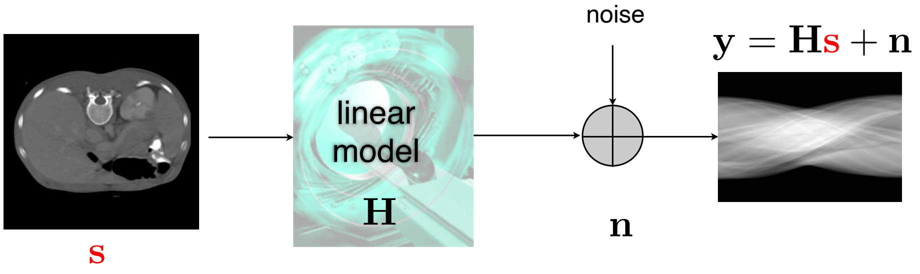
- **Introduction** ✔
- **Scientific context: Image reconstruction**
  - Classical image reconstruction
  - Compressed sensing and the sparsity revolution
  - Emergence of deep-CNN-based methods for image reconstruction
- **Can we trust CNN-based methods ?**
  - Dark sides of deep architectures
  - **Safeguards**: imposing **consistency** and **stability**
  - PnP framework with recurrent CNNs
- **Controlled design of nonlinearities**
  - Optimality of splines
  - Deep spline framework
- **Application to (stable) iterative image reconstruction**

AdG GlobalBioIm
(2016-2021)

AdG FunLearn
(2021-2026)

Scientific context: Image Reconstruction

- Inverse problem (typically ill-posed)



$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}$$

Goal: recover $\mathbf{s}$ from noisy measurements $\mathbf{y}$

- Classical paradigm: Formulation as an optimization problem

$$\mathbf{s}_{\mathrm{rec}} = \arg\min_{\mathbf{s}\in\mathbb{R}^N} \underbrace{\|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2}_{\text{data consistency}} + \underbrace{\lambda\|\mathbf{L}\mathbf{s}\|_p^p}_{\text{regularization}}, \quad p = 1, 2$$

Classical image reconstruction $(p = 2)$

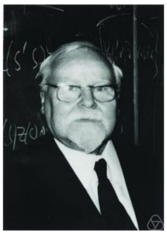
- Dealing with **ill-posed problems**: Tikhonov **regularization**

$\mathcal{R}(\mathbf{s}) = \|\mathbf{L}\mathbf{s}\|_2^2$: regularization (or smoothness) functional

$\mathbf{L}$: regularization operator (i.e., Gradient)

$$\min_{\mathbf{s}} \mathcal{R}(\mathbf{s}) \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2 \le \sigma^2$$

(consistency)



Andrey N. Tikhonov (1906-1993)

- Equivalent variational problem

$$\mathbf{s}^\star = \arg\min \underbrace{\|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2}_{\text{data consistency}} + \underbrace{\lambda\|\mathbf{L}\mathbf{s}\|_2^2}_{\text{regularization}}$$

Formal linear solution: $\mathbf{s} = (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{L}^T\mathbf{L})^{-1}\mathbf{H}^T\mathbf{y} = \mathbf{R}_\lambda \cdot \mathbf{y}$

Interpretation: “filtered” backprojection

Image reconstruction under sparsity constraints (CS) $p = 1$

- Convex optimization problem with non-smooth regularization (Donoho, Candes, 2006)

(1) $$\mathbf{s}_{\text{sparse}} = \arg\min_{\mathbf{s}} \left(\tfrac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2 + g(\mathbf{s})\right)$$ with $g(\mathbf{s}) = \lambda\|\mathbf{L}\mathbf{s}\|_{\ell_1}$ (regularization)

- Solution by forward-backward splitting (Combettes-Wajs, 2005)

Repeat

$\mathbf{z}^{(n)} = \mathbf{s}^{(n-1)} + \gamma\left(\mathbf{H}^\mathsf{T}\mathbf{y} - \mathbf{H}^\mathsf{T}\mathbf{H}\mathbf{s}^{(n-1)}\right)$ Linear step (**consistency** with imaging physics)

$\mathbf{s}^{(n)} = \text{prox}_{\gamma g}\left(\mathbf{z}^{(n)}\right)$ Proximal step (**regularization** = prior constraints)

until stop criterion

Guarantee of convergence: $\gamma \le \frac{2}{\lambda_{\max}(\mathbf{H}^\mathsf{T}\mathbf{H})}$

Proximal operator: $\text{prox}_g(\mathbf{z}) = \arg\min_{\mathbf{s}} \left(\tfrac{1}{2}\|\mathbf{z} - \mathbf{s}\|_2^2 + g(\mathbf{s})\right)$ (Moreau 1962)

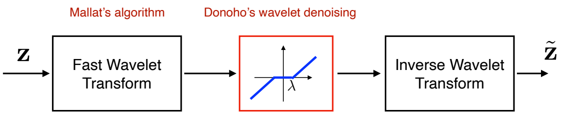
Interpretation: Same as (1) with $\mathbf{H} = \mathbf{I}$ $\Rightarrow$ “denoising” of current estimate $\mathbf{z}$

Efficient proximal denoising: wavelet-domain soft thresholding

- Regularization: Promote sparsity in an orthogonal wavelet basis

$g(\mathbf{s}) = \lambda \|\mathbf{W}^{\mathsf{T}}\mathbf{s}\|_{\ell_1}$ with $\mathbf{W}^{\mathsf{T}}\mathbf{W} = \mathbf{I}$ (Orthonormality)

Proximal step: $\tilde{\mathbf{z}} = \mathrm{prox}_g(\mathbf{z}) = \arg\min_{\mathbf{s}} \left(\frac{1}{2}\|\mathbf{z} - \mathbf{s}\|_2^2 + \lambda \|\mathbf{W}^{\mathsf{T}}\mathbf{s}\|_{\ell_1}\right)$



- Iterative Soft-Thresholding Algorithm (ISTA) (Figueiredo-Nowak 2003)

Repeat

$$\mathbf{z}^{(n)} = \mathbf{s}^{(n-1)} + \frac{1}{L}\left(\mathbf{H}^{\mathsf{T}}\mathbf{y} - \mathbf{H}^{\mathsf{T}}\mathbf{H}\mathbf{s}^{(n-1)}\right)$$

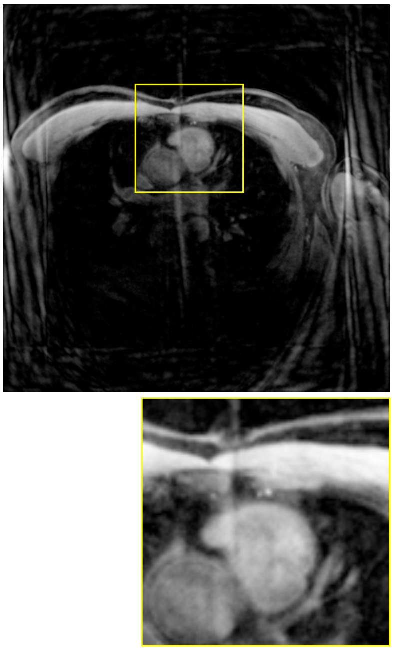
$$\mathbf{s}^{(n)} = \mathrm{WaveletThreshold}\left(\mathbf{z}^{(n)}; \lambda/L\right)$$

until stop criterion

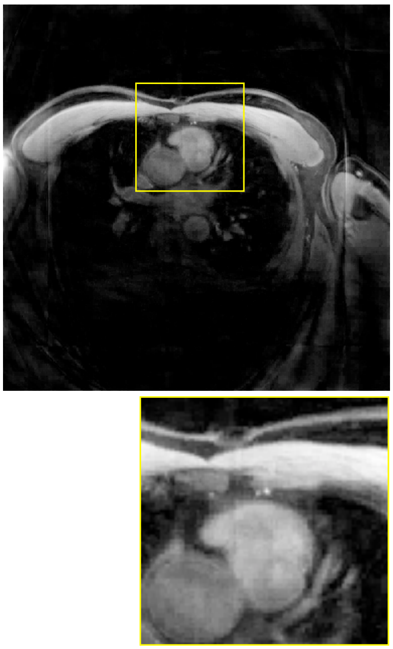
… and variants: WISTA, FISTA, …

ISMRM reconstruction challenge

L_2 regularization (Laplacian)



ℓ_1 wavelet regularization



WISTA

Collaboration with
Prof. Klass Prüssmann

ETH zürich

(Guerquin-Kern *IEEE TMI* 2011)

Compressed sensing: Applications in imaging

- Magnetic resonance imaging (MRI) (Lustig-Donoho, *Mag. Res. Im.* 2007)

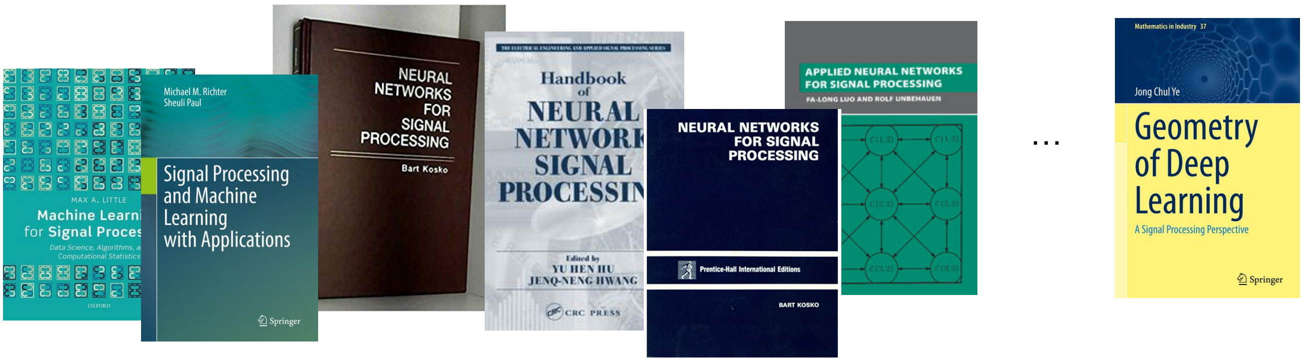
GE Healthcare PHILIPS SIEMENS

- Radio Interferometry (Wiaux, *Notic. R. Astro.* 2007)
- Teraherz Imaging (Chan, *Appl. Phys.* 2008)
- X-ray (interior) tomography (Wang, *Phys. Med. & Biol*; 2009)
- Digital holography (Brady, *Opt. Express* 2009; Marim 2010)
- Spectral-domain OCT (Liu, *Opt. Express* 2010)
- Coded-aperture spectral imaging (Arce, *IEEE Sig. Proc.* 2014)
- Localization microscopy (Zhu, *Nat. Meth.* 2012)
- Ultrafast photography (Gao, *Nature* 2014)

The (deep) learning (r)evolution in image processing

Special issues

Flurry of new textbooks on neural networks

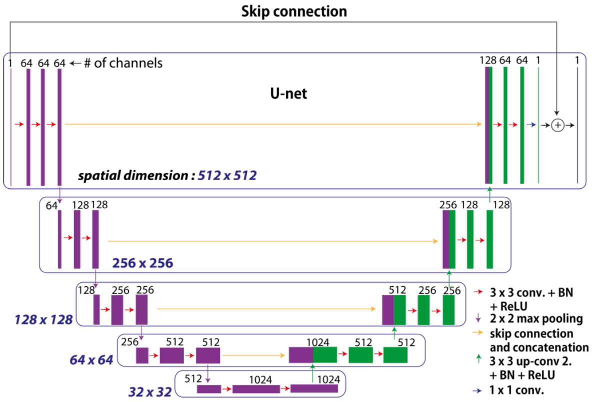

Machine Learning for Signal Processing
Signal Processing and Machine Learning with Applications
NEURAL NETWORKS FOR SIGNAL PROCESSING
Handbook of NEURAL NETWORK SIGNAL PROCESSING
NEURAL NETWORKS FOR SIGNAL PROCESSING
APPLIED NEURAL NETWORKS FOR SIGNAL PROCESSING
…
Geometry of Deep Learning

Appearance of Deep ConvNets

(Jin et al. 2016; Adler-Öktem 2017; Chen et al. 2017; ...)

(Jin et al., IEEE TIP 2017)

- CT reconstruction based on Deep ConvNets
  - Input: Sparse view FBP reconstruction
  - Training: Set of 500 high-quality full-view CT reconstructions
  - Architecture: U-Net with skip connection



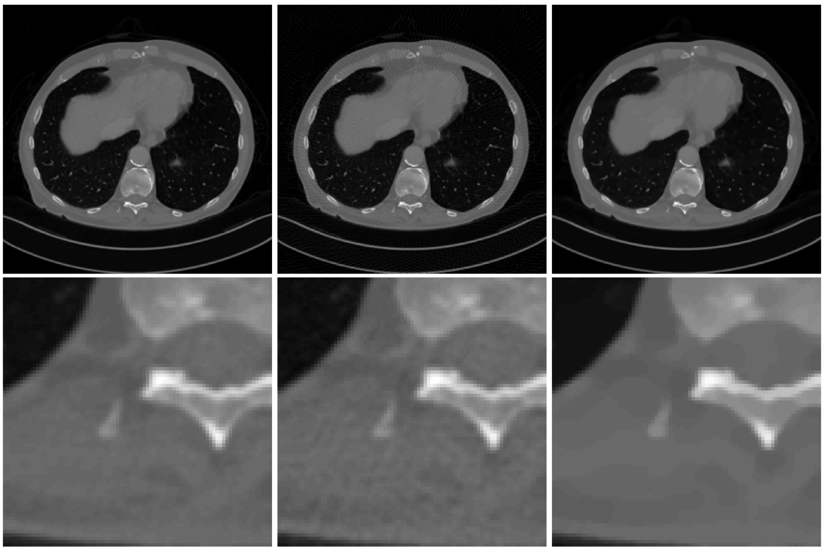
CT data

Dose reduction by 7: 143 views

Ground truth

FBP
SNR 24.06

TV
SNR 29.64

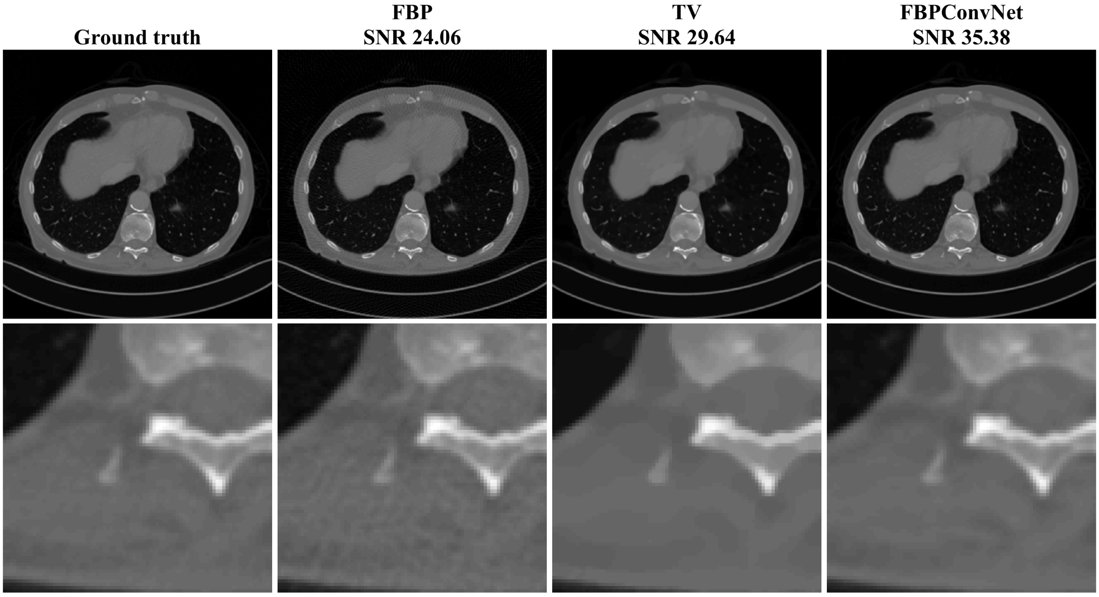


Reconstructed from
from 1000 views

MAYO CLINIC

CT data

Dose reduction by 7: 143 views



Reconstructed from
from 1000 views

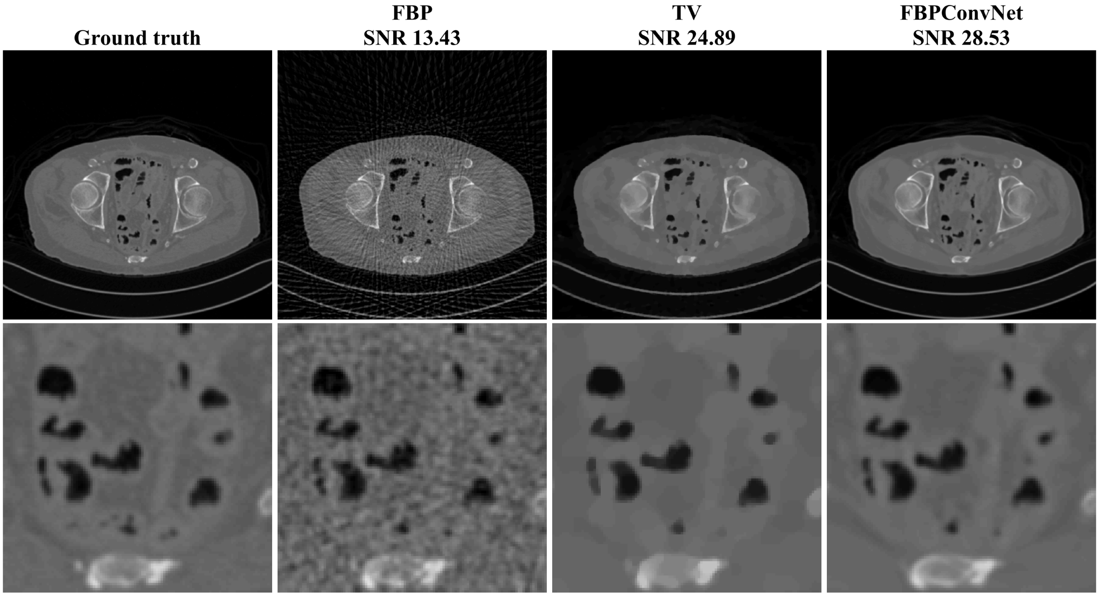
MAYO CLINIC

(Jin et al, *IEEE Trans. Im Proc.*, 2017)

IEEE
2019 Best Paper Award

CT data

Dose reduction by 20: 50 views



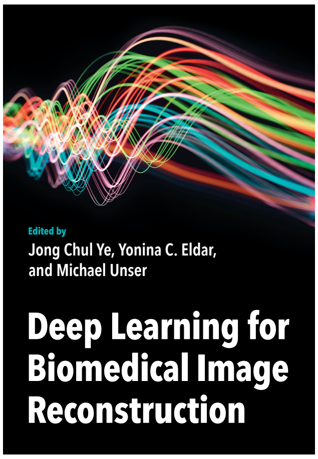
Reconstructed from
from 1000 views

MAYO CLINIC

(Jin et al., *IEEE Trans. Im Proc.*, 2017)

Deep CNNs for bioimage reconstruction images

- X-ray tomography	(Jin···Unser, *IEEE TIP* 2017) (Chen···Wang, *Biomed Opt. Exp* 2017)
- Magnetic resonance imaging (MRI)	(Hammernik···Pock, *Mag Res Med* 2018) (Tezcan···Konukoglu, *IEEE TMI* 2018)
- Dynamic MRI (cardial imaging)	(Schlemper···Rueckert, *IEEE TMI 2018*) (Hauptmann···Arridge, *Mag Res Med* 2019)
- 2D microscopy	(Rivenson···Ozcan, *Optica* 2017)
- 3D fluorescence microscocopy	(Weigert···Jug, Myers, *Nature Meth. 2018*)
- Super-resolution microscopy	(Nehme···Shechtman, *Optica* 2018)
- Diffraction tomography	(Sun···Kamilov, *Optics Express* 2018)
- Ultrasound	(Yoon···Ye, *IEEE TMI* 2019)


Edited by
Jong Chul Ye, Yonina C. Eldar,
and Michael Unser

Deep Learning for Biomedical Image Reconstruction

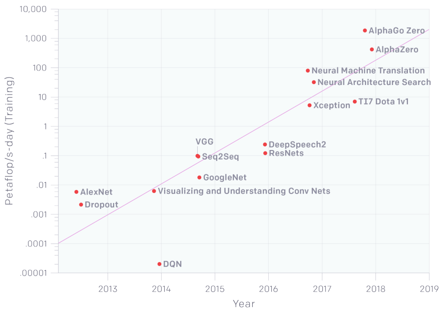
OUTLINE

- Introduction ✔
- Scientific context: Image reconstruction ✔
- **Can we trust CNN-based methods ?**
  - The dark side of deep architectures
  - **Safeguards**: imposing consistency and stability
  - PnP framework with recurrent CNNs
- Controlled design of nonlinearities
- Application to (stable) iterative image reconstruction

But CNN-based methods also have their weaknesses

- They require **lots of training data**
  - Medical imaging: limited access to patient data
  - Lack of gold standards (except for compressed sensing scenarios)
  - Training for (3D) medical imaging is **extremely computer intensive**

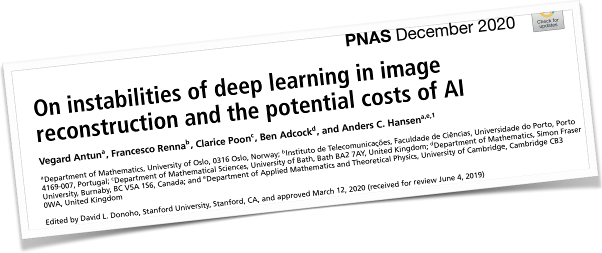
AI Moore's law: doubling every 3.4 month !



source: openAI

- They are **hard to tune**
  - Many design parameters: depth, width, number of channels
  - Use of ad hoc modules: batch normalization

- They **lack robustness**
  - Adversarial attacks
  - Unpredictable results



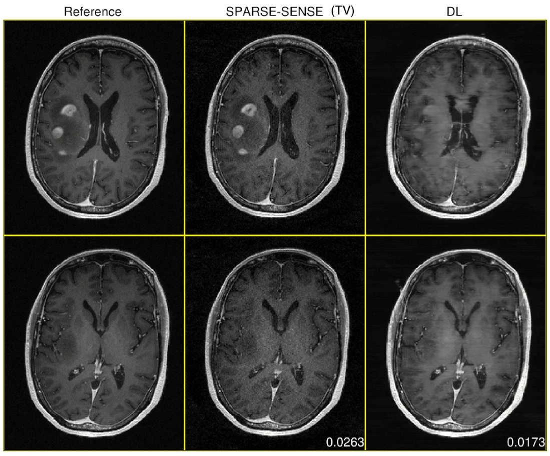
PNAS December 2020

On instabilities of deep learning in image reconstruction and the potential costs of AI

Vegard Antun[a], Francesco Renna[b], Clarice Poon[c], Ben Adcock[d], and Anders C. Hansen[a,e,1]

[a]Department of Mathematics, University of Oslo, 0316 Oslo, Norway; [b]Instituto de Telecomunicações, Faculdade de Ciências, Universidade do Porto, Porto 4169-007, Portugal; [c]Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, United Kingdom; [d]Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada; and [e]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, United Kingdom

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved March 12, 2020 (received for review June 4, 2019)



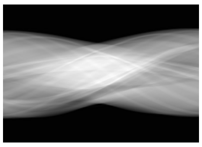
(variant of FBPConvNet)

Figure 3: Reconstructions in a case of anaplastic astrocytoma, a rare malignant brain tumor. SPARSE-SENSE and DL reconstructions are from the same 4x-accelerated retrospectively undersampled acquisition. DL achieves lower whole-volume MAE than SPARSE-SENSE, but fails to properly reconstruct regions near the tumor.

G. Nataraj and R. Otazo. "Investigating robustness to unseen pathologies in model-free deep multicoil reconstruction."
ISMRM 2020 Workshop on Data Sampling & Image Reconstruction

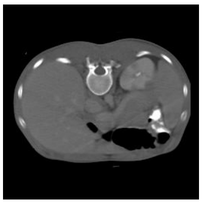
Mathematical safeguards

Forward imaging model: $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}_{\mathrm{noise}}$



Data

Image reconstruction algorithm: $\tilde{\mathbf{x}} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{y})$



Reconstruction

- Consistency of reconstruction

$\|\mathbf{y} - \mathbf{H}\tilde{\mathbf{x}}\| = \|\mathbf{y} - \mathbf{H}\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{y})\| \leq \epsilon$ for some suitable ϵ

- Stability of reconstruction algorithm

$\|\tilde{\mathbf{x}}_2 - \tilde{\mathbf{x}}_1\| = \|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{y}_2) - \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{y}_1)\| \leq L\, \|\mathbf{y}_2 - \mathbf{y}_1\|,$ for all $\mathbf{y}_2, \mathbf{y}_1 \in \Omega \subseteq \mathbb{R}^M$

with $L = \mathrm{Lip}(\mathbf{f}_{\boldsymbol{\theta}})$ reasonably small

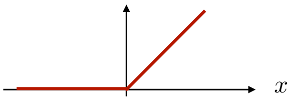
(del Aguila Pla IEEE TCI, 2023)

Lipschitz constant of primary modules

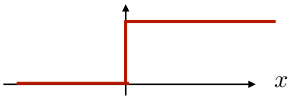
- Pointwise nonlinearity

$\sigma : \mathbb{R} \to \mathbb{R}$ where σ is differentiable

$$\mathrm{Lip}(\sigma) = \sup_{x\in\mathbb{R}} \left| \frac{\mathrm{d}\sigma(x)}{\mathrm{d}x} \right| = \|\sigma'\|_{L_\infty}$$ (cf. Mean Value Theorem)



Heaviside



Example: $\mathrm{Lip}(\mathrm{ReLU}) = \sup_{x\in\mathbb{R}} |u(x)| = 1$

- Linear (resp. affine) transform

$\mathrm{T}_{\mathrm{lin}} : \mathbb{R}^M \to \mathbb{R}^N$ with $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ (linear)

or $\mathbf{x} \mapsto \mathbf{A}\mathbf{x} + \mathbf{b}$ (affine) where $\mathbf{A} \in \mathbb{R}^{M\times N}, \mathbf{b} \in \mathbb{R}^M$

$$\mathrm{Lip}(\mathrm{T}_{\mathrm{lin}}) = \sup_{\|\mathbf{x}\|_2 \leq 1} \|\mathbf{A}\mathbf{x}\|_2 = \rho(\mathbf{A})$$ (spectral norm = largest singular value of $\mathbf{A}$)

- Composition

$$\xrightarrow{} \boxed{\mathrm{T}_1} \xrightarrow{} \boxed{\mathrm{T}_2} \xrightarrow{}$$

$\mathrm{Lip}(\mathrm{T}_1) = L_1$ & $\mathrm{Lip}(\mathrm{T}_2) = L_2 \quad \Rightarrow \quad \mathrm{Lip}(\mathrm{T}_2 \circ \mathrm{T}_1) \leq L_2 L_1$

Consistency via PnP variant of iterative reconstruction

Schematic structure of iterative reconstruction algorithm : $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left(\frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + g(\mathbf{x}) \right)$

```
N_iter
  Repeat
      $\mathbf{z}^{(n)} = \mathbf{x}^{(n-1)} + \alpha \left( \mathbf{H}^{\mathsf{T}}\mathbf{y} - \mathbf{H}^{\mathsf{T}}\mathbf{H}\mathbf{x}^{(n-1)} \right)$      Linear step (consistency with imaging physics)
      $\mathbf{x}^{(n)} = \mathrm{prox}_{\alpha g}\left(\mathbf{z}^{(n)}\right)$      Proximal or "denoising" step (regularization)
  until stop criterion
```

Proximal operator: $\mathrm{prox}_g(\mathbf{z}) = \arg\min_{\mathbf{x}} \left(\frac{1}{2}\|\mathbf{z} - \mathbf{x}\|^2 + g(\mathbf{x}) \right)$

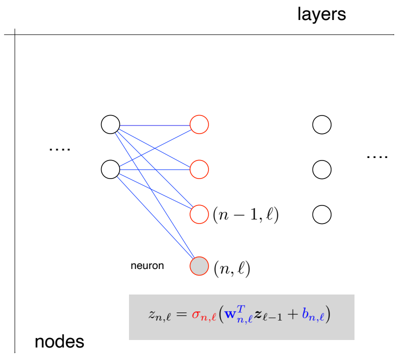
Plug-and-Play variant (Venkatakrishnan-Bouman 2013)

```
N_iter
  Repeat
      $\mathbf{z}^{(n)} = \mathbf{x}^{(n-1)} + \alpha \left( \mathbf{H}^{\mathsf{T}}\mathbf{y} - \mathbf{H}^{\mathsf{T}}\mathbf{H}\mathbf{x}^{(n-1)} \right)$      Linear step (consistency with imaging physics)
      $\mathbf{x}^{(n)} = \left( (1-\beta)\mathrm{Id} + \beta \boldsymbol{f}_{\boldsymbol{\theta}} \right)\left(\mathbf{z}^{(n)}\right)$      Suitable nonlinear map (e.g., CNN)
  until stop criterion
```

Requirement for convergence: $\|\boldsymbol{f}_{\boldsymbol{\theta}}\|_{\mathrm{Lip}} \leq 1$ (Non-expansive operator) (Bauschke-Combettes 2017, Hertrich et al. 2021)

Neural nets with free-form activations and stability control

- Layers: $\ell = 1, \dots, L$
- Deep structure descriptor: $(N_0, N_1, \cdots, N_L)$
- Neuron or node index: (n, ℓ), $n = 1, \cdots, N_\ell$
- Activation function $\sigma_{n,\ell} : \mathbb{R} \to \mathbb{R}$ (free-form)
- Linear step: $\mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{f}_\ell : \boldsymbol{x} \mapsto \boldsymbol{f}_\ell(\boldsymbol{x}) = \mathbf{W}_\ell \boldsymbol{x} + \mathbf{b}_\ell$
- Nonlinear step: $\mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{\sigma}_\ell : \boldsymbol{x} \mapsto \boldsymbol{\sigma}_\ell(\boldsymbol{x}) = \left(\sigma_{n,\ell}(x_1), \dots, \sigma_{N_\ell,\ell}(x_{N_\ell}) \right)$



$z_{n,\ell} = \sigma_{n,\ell}\left(\mathbf{w}_{n,\ell}^{T} \boldsymbol{z}_{\ell-1} + b_{n,\ell}\right)$

$$\boldsymbol{f}_{\mathrm{deep}}(\boldsymbol{x}) = \left(\boldsymbol{\sigma}_L \circ \boldsymbol{f}_L \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \circ \boldsymbol{\sigma}_2 \circ \boldsymbol{f}_2 \circ \boldsymbol{\sigma}_1 \circ \boldsymbol{f}_1 \right)(\boldsymbol{x})$$

Joint learning / training

Stability control: $\|\boldsymbol{f}_{\mathrm{deep}}\|_{\mathrm{Lip}} \leq \prod_{\ell=1}^{L} \underbrace{\|\boldsymbol{\sigma}_\ell\|_{\mathrm{Lip}}}_{1} \underbrace{\rho(\mathbf{W}_\ell)}_{1} = 1$

Lip-1 splines

spectral normalization vs. Parseval frame

OUTLINE

- Introduction ✔
- Scientific context: Image reconstruction ✔
- Can we trust CNN-based methods? ✔
  - **Safeguards**: imposing consistency and stability
- **Controlled design of nonlinearities**
  - Optimality of splines
  - Deep spline framework
- Application to (stable) iterative image reconstruction

Learning activation functions / pointwise nonlinearities

Finding the "optimal" pointwise nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$

Infinite-dimensional optimization problem is that is inherently ill-posed

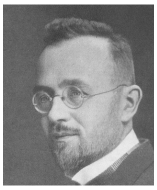
- Incorporating a **regularization**
  - Should not penalize simple solutions (e.g., identity or linear scaling)
  - Should impose differentiability (for DNN to be trainable via backpropagation)
  - Should favour simplest CPWL solutions; i.e., with "sparse 2nd derivatives"

$\Rightarrow$ minimizing/constraining $\mathrm{TV}^{(2)}(\sigma) \triangleq \|\mathrm{D}^2\sigma\|_{\mathcal{M}}$ (Second-order total-variation)

- Controlling **stability**: $\mathrm{Lip}(\sigma) \triangleq \sup_{x\in\mathbb{R}} |\mathrm{D}\sigma(x)| \leq 1$

- Search space: $\mathrm{BV}^{(2)}(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} : \|\mathrm{D}^2 f\|_{\mathcal{M}} < \infty\} \subset \mathrm{Lip}(\mathbb{R})$

Proper continuous counterpart of ℓ_1-norm



Johann Radon (1887-1956)

- Dual definition of ℓ_1-norm (in finite dimensions only)

$$\|\boldsymbol{f}\|_{\ell_1} = \sum_{n=1}^{N} |f_n| = \sup_{\boldsymbol{u}\in\mathbb{R}^N:\ \|\boldsymbol{u}\|_\infty \le 1} \langle \boldsymbol{f}, \boldsymbol{u}\rangle$$

- Space $C_0(\mathbb{R}^d)$ of functions on $\mathbb{R}^d$ that are continuous, bounded, and decaying at infinity

$$C_0(\mathbb{R}^d) = \overline{(\mathcal{S}(\mathbb{R}^d), \|\cdot\|_{L_\infty})} \ \subset\ L_\infty(\mathbb{R}^d)$$

- Space of **bounded Radon measures** on $\mathbb{R}^d$

$$\mathcal{M}(\mathbb{R}^d) = \big(C_0(\mathbb{R}^d)\big)' = \{f \in \mathcal{S}'(\mathbb{R}^d) : \|f\|_{\mathcal{M}} \stackrel{\Delta}{=} \sup_{\varphi\in\mathcal{S}(\mathbb{R}^d):\ \|\varphi\|_\infty\le 1} \langle f, \varphi\rangle < +\infty\}$$

- **Superset** of $L_1(\mathbb{R}^d)$

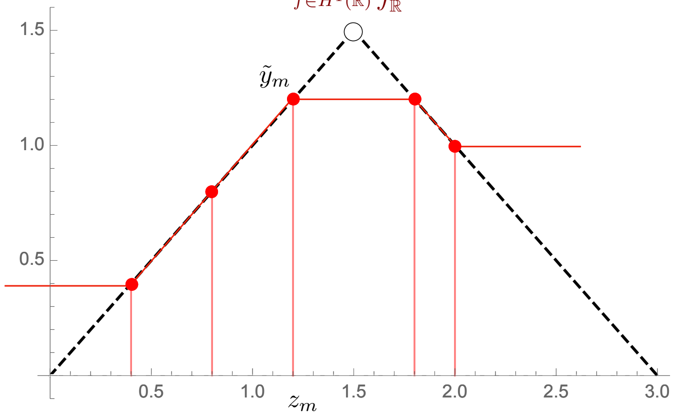
  $\forall f \in L_1(\mathbb{R}^d):\quad \|f\|_{\mathcal{M}} = \|f\|_{L_1} \quad \Rightarrow \quad L_1(\mathbb{R}^d) \subset \mathcal{M}(\mathbb{R}^d)$

- **Extreme points** of unit ball in $\mathcal{M}(\mathbb{R}^d)$: $\quad e_k = \pm\delta(\cdot - \boldsymbol{\tau}_k)$ with $\boldsymbol{\tau}_k \in \mathbb{R}^d$

Comparison of linear interpolators

$$\arg \min_{f\in H^1(\mathbb{R})} \int_{\mathbb{R}} |\mathrm{D}f(x)|^2 \mathrm{d}x \quad \text{s.t.} \quad f(x_m) = y_m,\ m = 1, \dots, M$$

(de Boor 1966)



$$\arg \min_{f\in \mathrm{BV}^{(2)}(\mathbb{R})} \|\mathrm{D}^2 f\|_{\mathcal{M}} \quad \text{s.t.} \quad f(x_m) = y_m,\ \ m = 1, \dots, M$$

(Unser JMLR 2019; Lemma 2)

Optimality of splines: TV$^{(2)}$ regularization with slope constraints

Training data: $(x_m, y_m) \in \mathbb{R} \times \mathbb{R}, \quad m = 1, \dots, M$

Generic loss functional $E : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ (strictly convex)

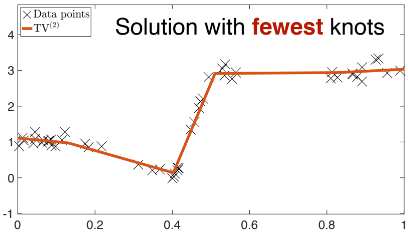
Slope parameters: $s_{\min} < s_{\max}$

(TV2-SC)
$$S = \arg \min_{f \in \mathrm{BV}^{(2)}(\mathbb{R})} \left(\sum_{m=1}^{M} \mathrm{E}(f(x_m), y_m) + \lambda \mathrm{TV}^{(2)}(f) \right),$$
$$\text{s.t.} \quad s_{\min} \le f'(x) \le s_{\max}, \quad \forall x \in \mathbb{R}$$

Theorem (new improved: for Stéphane Mallat's birthday - April 2023)

The solution set of (TV2-SC) is a non-empty, weak*-compact subset of $\mathrm{BV}^{(2)}(\mathbb{R})$, and **all its extreme points** are **adaptive piecewise-linear splines** with a most $(M-2)$ knots.

Sparsest spline solution is identifiable using a variant of Debarre's algorithm.



(Debarre JCAM 2022)

$(s_{\min}, s_{\max}) = \mathbb{R}$ (unconstrained)

- Special cases of $(s_{\min}, s_{\max})$
  - $(-1, 1)$: Lipschitz-1 splines (Aziznejad, IEEE OJSP 2022)
  - $(0, 1)$: firmly non-expansive = prox of a convex potential
  - $(0, +\infty)$: monotone splines = derivative of a convex potential — invertible function
  - $(-\rho, +\infty)$ with $0 < \rho$ (small): weakly-monotone splines = derivative of a ρ-weakly-convex potential

Representer theorem for stable, free-form deep neural networks

Theorem (Optimality of Lipschitz-1 deep spline networks)

- neural network $\mathbf{f} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ with **deep structure** $(N_0, N_1, \dots, N_L)$

  $\boldsymbol{x} \mapsto \boldsymbol{f}_{\text{deep}}(\boldsymbol{x}) = (\boldsymbol{\sigma}_L \circ \boldsymbol{f}_L \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \circ \boldsymbol{f}_2 \circ \boldsymbol{\sigma}_1 \circ \boldsymbol{f}_1)(\boldsymbol{x})$
- linear transformations $\boldsymbol{f}_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, \boldsymbol{x} \mapsto \mathbf{W}_\ell \boldsymbol{x}$ with $\mathbf{W}_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$
- **free-form** activations $\boldsymbol{\sigma}_\ell = (\sigma_{1,\ell}, \dots, \sigma_{N_\ell,\ell}) : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$ with $\sigma_{1,\ell}, \dots, \sigma_{N_\ell,\ell} \in \mathrm{BV}^{(2)}(\mathbb{R})$

Given a series data points $(\boldsymbol{x}_m, \boldsymbol{y}_m)$ $m = 1, \dots, M$, we then define the training problem

$$\arg \min_{(\mathbf{W}_\ell),(\sigma_{n,\ell} \in \mathrm{BV}^{(2)}(\mathbb{R}))} \left(\sum_{m=1}^{M} E(\boldsymbol{y}_m, \boldsymbol{f}_{\text{deep}}(\boldsymbol{x}_m)) + \lambda \sum_{\ell=1}^{L} \sum_{n=1}^{N_\ell} \mathrm{TV}^{(2)}(\sigma_{n,\ell}) \right)$$
$$\text{s.t.} \quad \mathrm{Lip}(\sigma_{n,\ell}), \rho(\mathbf{W}_\ell) \le 1, \quad (n = 1, \dots, N_\ell, \ \ell = 1, \cdots, L) \qquad (1)$$

$$\Rightarrow \mathrm{Lip}(\boldsymbol{f}_{\text{deep}}) \le 1$$

where $E : \mathbb{R}^{N_L} \times \mathbb{R}^{N_L} \to \mathbb{R}^+$ is an arbitrary convex loss function.

The solution of (1) exists and is achieved by a **deep spline network** with activations of the form

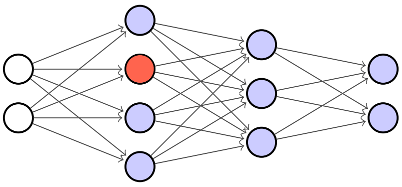
$$\sigma_{n,\ell}(x) = b_{1,n,\ell} + b_{2,n,\ell} x + \sum_{k=1}^{K_{n,\ell}} a_{k,n,\ell}(x - \tau_{k,n,\ell})_+,$$

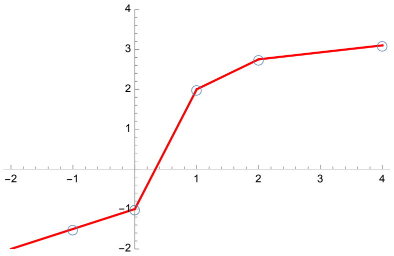
with adaptive parameters $K_{n,\ell} \le M - 2$, $\tau_{1,n,\ell}, \dots, \tau_{K_{n,\ell},n,\ell} \in \mathbb{R}$, and $b_{1,n,\ell}, b_{2,n,\ell}, a_{1,n,\ell}, \dots, a_{K_{n,\ell},n,\ell} \in \mathbb{R}$.

Precursor without stability: (Unser, JMLR 2019)

Outcome of representer theorem

$$\sigma_{n,\ell}(x) = b_{1,n,\ell} + b_{2,n,\ell}x + \sum_{k=1}^{K_{n,\ell}} a_{k,n,\ell}(x - \tau_{k,n,\ell})_+,$$





Each neuron $\big($fixed index $(n,\ell)\big)$ is characterized by

- its number $K = K_{n,\ell} \geq 0$ of knots (ideally, much smaller than M);
- the locations $\{\tau_k = \tau_{k,n,\ell}\}_{k=1}^{K}$ of these knots;
- the expansion coefficients $\mathbf{b}_{n,\ell} = (b_{1,n,\ell}, b_{2,n,\ell}) \in \mathbb{R}^2$, $\mathbf{a}_{n,\ell} = (a_{1,n,\ell}, \dots, a_{K,n,\ell}) \in \mathbb{R}^K$.

These parameters (including the number of knots) are **data-dependent** and **must be adjusted** (automatically) **during training**.

■ Link with ℓ_1 minimization techniques

$$\mathrm{TV}^{(2)}(\sigma_{n,\ell}) = \sum_{k=1}^{K_{n,\ell}} |a_{k,n,\ell}| = \|\mathbf{a}_{n,\ell}\|_1 \qquad \text{and} \qquad \mathrm{Lip}(\sigma_{n,\ell}) = \sup_{K \in \{1,\dots,K_{n,\ell}\}} \left| \sum_{k=1}^{K} a_{k,n,\ell} \right|$$

How to effectively train deep splines ?

Stochastic **gradient descent** (the difficult part being to optimize the knot locations)

Workaround: Fixed set of knots on a **grid**—rely on ℓ_1-minimization to suppress the unnecessary ones

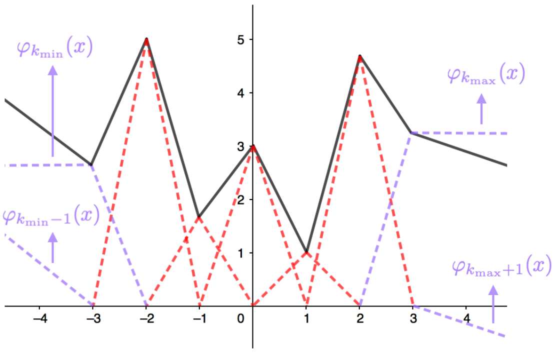
■ Gridded ReLU representation

$$\sigma(x) = b_0 + b_1 x + \sum_{k=k_{\min}}^{k_{\max}} a_k (x - kT)_+$$

■ B-spline representation

$$\sigma(x) = \sum_{k=k_{\min}-1}^{k_{\max}+1} c_k \varphi_k\left(\frac{x}{T}\right)$$

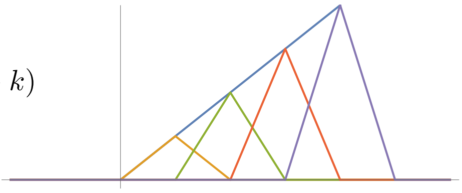
where $\varphi_k(x) = \mathrm{tri}(x-k),$ for $k_{\min} < k < k_{max}$



Equivalence between ReLU and B-spline representations

Simplified cardinal spline setting with $T = 1$ and $\varphi_k(x) = \mathrm{tri}(x-k), k \in \mathbb{Z}$

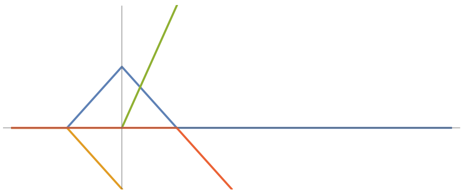
- Expressivity of triangular B-spline basis



Polynomials: $$1 = \sum_{k\in\mathbb{Z}} \mathrm{tri}(x-k), \qquad x = \sum_{k\in\mathbb{Z}} k\,\mathrm{tri}(x-k)$$

Gridded ReLUs: $$(x-k_0)_+ = \sum_{k=k_0}^{+\infty} (k-k_0)\mathrm{tri}(x-k)$$

- From ReLUs to B-splines



$$\mathrm{tri}(x) = -1(x+1)_+ + 2(x)_+ - 1(x-1)_+$$

- Second total variation

$$\sigma(x) = \sum_{k\in\mathbb{Z}} c[k]\mathrm{tri}(x-k) \quad \Rightarrow \quad \mathrm{TV}^{(2)}(\sigma) = \|d_2 * c\|_{\ell_1}$$

2nd difference filter: $d_2[\cdot] = (-1, 2, -1)$

B-spline basis—complexity is independent of grid size !

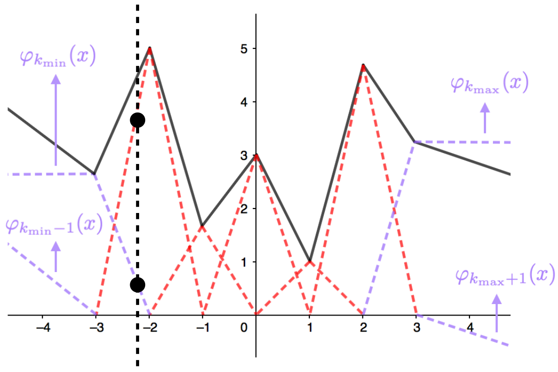


TABLE IV: B-splines *vs.* gridded ReLUs *vs.* APLUs

Architecture, Nb. coefficients	Memory (megabytes)	Time per epoch (seconds)
B-splines, $K = 9$	1132	44.92
B-splines, $K = 29$	1133	41.89
B-splines, $K = 499$	1299	41.19
Gridded ReLUs, $K = 9$	3313	49.86
Gridded ReLUs, $K = 29$	9616	81.21
APLUs, $K = 9$	3316	49.72
APLUs, $K = 29$	9618	87.34

For the gridded ReLU and APLU networks, the maximum number of knots allowed by the GPU memory is 31.

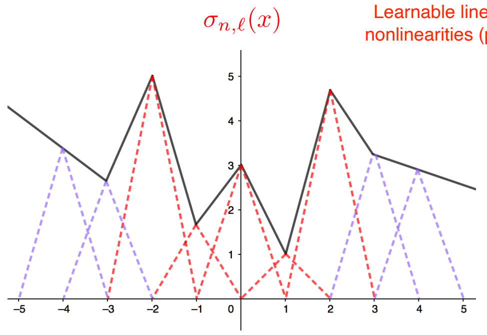
Explanation: only **two** active basis functions per data point

Implementation: Lip-1 spline CNN (trained for denoising)

$$\boldsymbol{f}_{\mathrm{deep}}(\boldsymbol{x}) = (\boldsymbol{\sigma}_L \circ \boldsymbol{f}_L \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \circ \boldsymbol{\sigma}_2 \circ \boldsymbol{f}_2 \circ \boldsymbol{\sigma}_1 \circ \boldsymbol{f}_1)(\boldsymbol{x})$$

Learnable linear spline nonlinearities (pointwise)

Convolutional layer



Knot spacing: T, Number of knots: K

- Linear B-spline basis
  - Compact support
  - Efficient forward & backward pass
  - Easy to compute Lipschitz constant (max. absolute derivative)

(Bohra et al. *IEEE Open JSP* 2020)

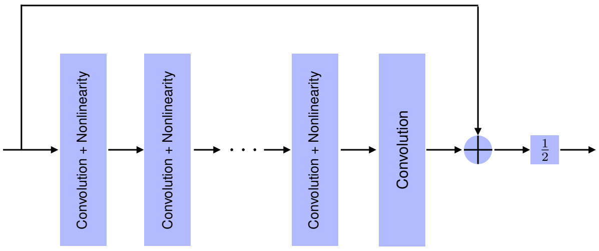
- Constrain Lipschitz constant of each layer to be no greater than one
  - Convolutional layer: Lip-1 projector (spectral normalization vs. Parseval frame)
  - Linear spline layer: Lip-1 spline projector (clipping of finite difference)

(Ducotterd et al. ArXiv 2022)

PnP image reconstruction: Experimental set-up

- Training of Gaussian denoiser
  - 240K examples of 40×40 patches from BSD500 dataset
  - Additive Gaussian noise with $\sigma = 5/255$
  - 3×3 convolution kernels, 32 channels
  - Deep spline activations with $T = 0.1$, $K = 51$
  - Number of layers = 3, 5, 7, 9
- Compressed sensing MRI
  - 256×256 ground-truth images
  - Subsampling ratio = 0.3
  - Gaussian additive noise with $\sigma = 10/255$
  - Number of layers of denoising CNN = 5

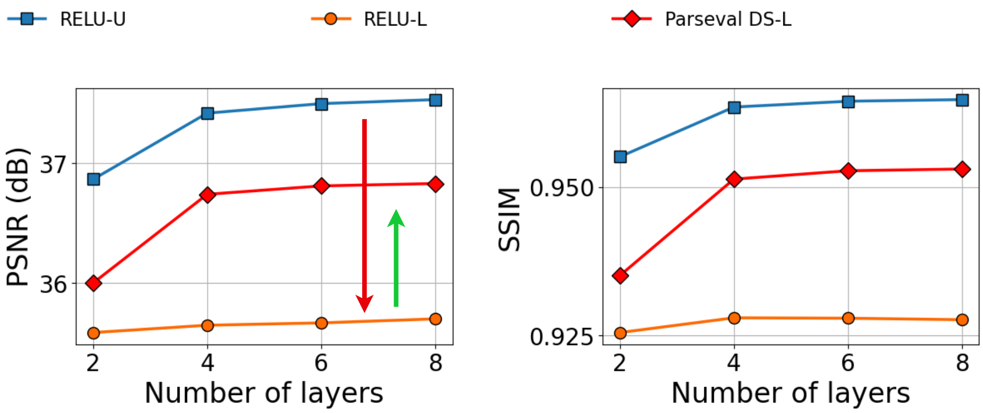
Deep-spline denoising Dn-CNN



Convolution + Nonlinearity
Convolution + Nonlinearity
· · ·
Convolution + Nonlinearity
Convolution
$\frac{1}{2}$

Learned Lip-1 filters = Parseval frames

Results: Gaussian denoising with Parseval frames

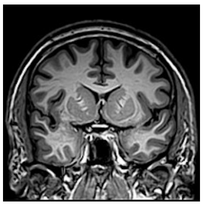


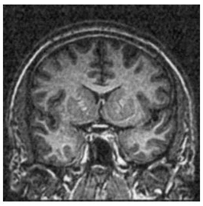
- ↓ Drop in performance for constrained ReLU nets
- ↑ DS-L performs better than ReLU-L even with fewer parameters

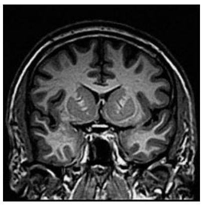
Compressed Sensing MRI

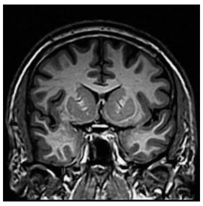
Subsampling mask	Random		Radial		Cartesian	
Image type	Brain	Bust	Brain	Bust	Brain	Bust
Zero-filling	23.72	25.88	22.99	23.92	21.34	23.03
ReLU-L	30.70	30.59	29.60	30.09	23.70	26.87
Parseval DS-L	33.19	33.88	31.68	33.15	24.97	28.68

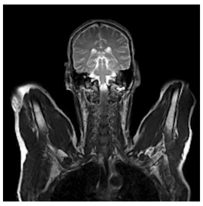
Random sampling pattern

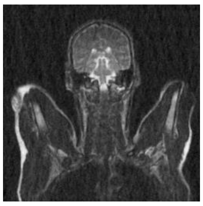

Ground truth

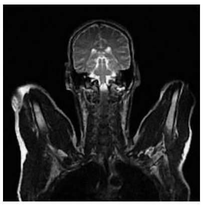

Least squares reconstruction (Zero-filled)

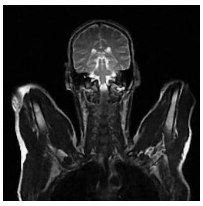

ReLU-L


Parseval DS-L


Ground truth


Least squares reconstruction (Zero-filled)


ReLU-L


Parseval DS-L

WCRR variant: Learnable Weakly-Convex Ridge Regularizer

$$\min_{\mathbf{x}\in\mathbb{R}^N}\left(\frac{1}{2}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2+\sum_{i=1}^{I_{\text{chan}}}\langle\mathbf{1},\boldsymbol{\Phi}_i(\mathbf{W}_i\mathbf{x})\rangle\right)$$

Weakly-convex extension of FoE (Chen-Pock 2014)

- System matrix: $\mathbf{H}\in\mathbb{R}^{M\times N}$
- Learnable filters (CNN) : $\mathbf{W}_i\in\mathbb{R}^{N\times N},\quad i=1,\dots,I_{\text{chan}}$
- Shared free-form potentials : $\boldsymbol{\Phi}_i(\mathbf{u})=\big(\Phi_i(u_1),\dots,\Phi_i(u_N)\big)$ with $\Phi_i(u)=\int_{-\infty}^{u}\phi_i(x)\mathrm{d}x$

Iterative reconstruction

Recurrent neural network (steepest descent)

$$\mathbf{x}^{(n+1)}=\mathbf{x}^{(n)}-\alpha\left(\sum_{i=1}^{I_{\text{chan}}}\mathbf{W}_i^{\mathsf{T}}\boldsymbol{\phi}_i(\mathbf{W}_i\mathbf{x}^{(n)})+\mathbf{H}^{\mathsf{T}}\big(\mathbf{H}\mathbf{x}^{(n)}-\mathbf{y}\big)\right)\quad\text{with}\quad\boldsymbol{\phi}_i=\boldsymbol{\Phi}_i'$$

Training on denoising problem

- Parametrization of the slope: $\phi_i=\Phi_i':\mathbb{R}\to\mathbb{R}$
  s.t. weak-monotonicity constraint and penalty on $\mathrm{TV}^{(2)}(\phi_i)$ (sparsity) $\Rightarrow$ **linear splines**
- Deep equilibrium training of variational denoiser where the ϕ_i are expanded in a B-spline basis.

Table 4.1

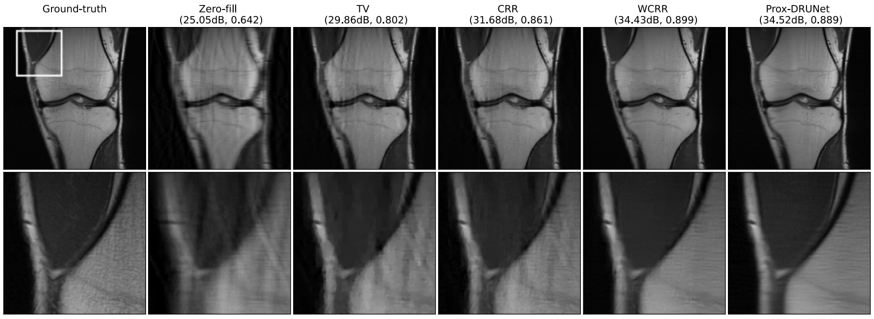
PSNR and SSIM values for both reconstruction experiments.

Metric	PSNR	SSIM
Zero-fill	27.92	0.711
TV[5]	32.03	0.7922
CRR-NN [19]	33.14	0.842
WCRR-NN	34.55	0.858
Prox-DRUNet [23]	35.09	0.864

(a) MRI

Metric	PSNR	SSIM	Param.
TV	31.57	0.852	1
ACR [37]	31.58	0.848	$6\cdot10^5$
CRR-NN	32.87	0.862	$5\cdot10^3$
AR [34]	33.62	0.875	$2\cdot10^7$
WCRR-NN	34.06	0.895	$2\cdot10^4$
Prox-DRUNet	34.20	0.901	$2\cdot10^7$

(b) CT



but, PSNR (or SSIM) is not the whole story

Theoretical guarantees : convergence, consistency, stability

convex handcrafted | convex learned | weakly convex learned | “state-of-the-art” CNN

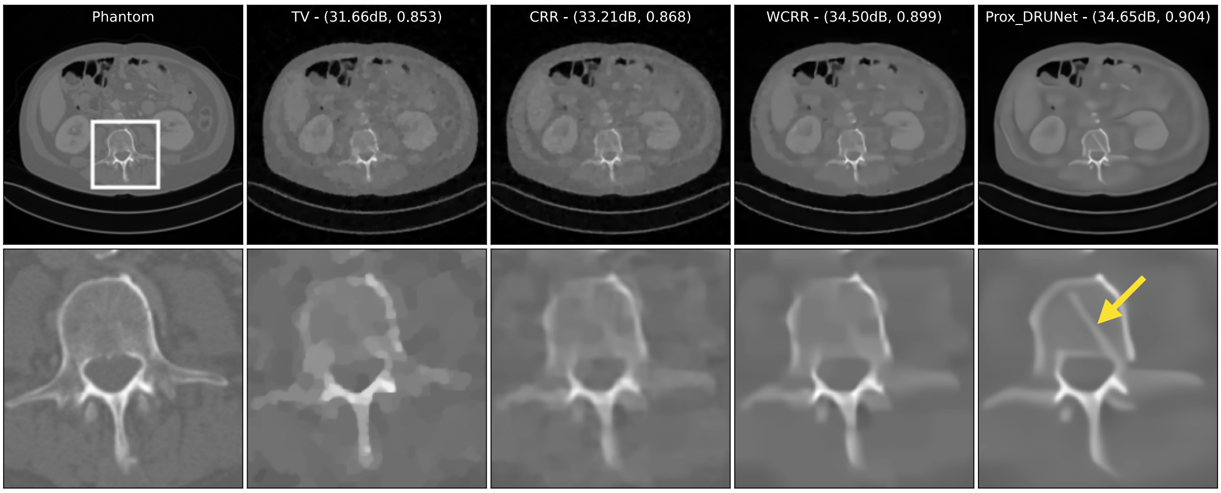
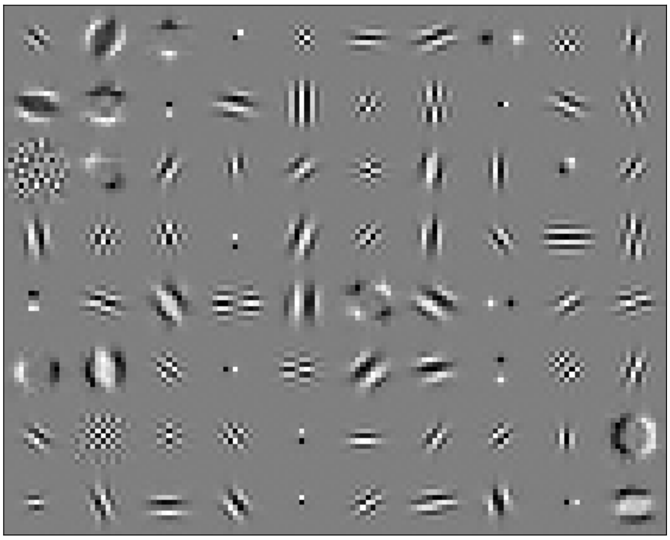
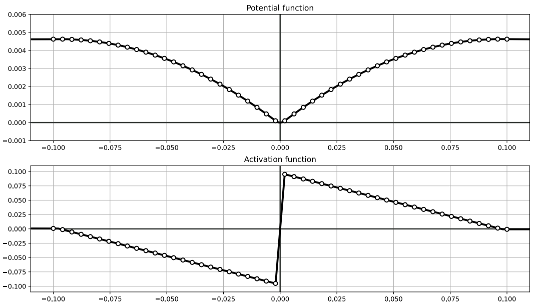


Figure 4.2. *Reconstructions for the sparse-view CT experiment. The reported metrics are PSNR and SSIM.*

Learned filters and nonlinearities

80 channels





Nonlinearities are shared up to a channel-wise scaling factor

Deep spline framework for learning nonlinearities

https://github.com/Biomedical-Imaging-Group/**DeepSplines**

- Versatibility
  - Lip-1 activations
  - Gradient of a (weakly-)convex potential
  - Proximity operator of a (weakly-)convex potential
  - Components of recurrent networks via deep equilibrium

- Typical usage
  - Revamping of traditional architectures (spirit of unrolling)
  - Refinement of not-so-deep architectures
  - Incorporation of stability constraints

- Quest for simplicity/interpretatibility
  - Ability to suppress unnecessary linear layers (via skip connection : $b_1 + b_2 x$)
  - Sharing a nonlinearity (up to an individual scaling factor)
  - Determination of the sparsest solution via the Debarre algorithm
  - Efficient encoding via non-uniform B-splines (during inference)

References

- Foundations
  - M.T. McCann, M. Unser, **Biomedical Image Reconstruction: From the Foundations to Deep Neural Networks**, *Foundations and Trends in Signal Processing*, vol. 13, no. 3, pp. 280-359, December 2019.
  - M. Unser, “A Unifying Representer Theorem for Inverse Problems and Machine Learning,” *Foundations of Computational Mathematics*, 2020. https://doi.org/10.1007/s10208-020-09472-x
- Algorithms and imaging applications
  - M. Guerquin-Kern, M. Häberlin, K.P. Pruessmann, M. Unser, ”A Fast Wavelet-Based Reconstruction Method for Magnetic Resonance Imaging,” *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1649-1660, 2011.
  - K.H. Jin, M.T. McCann, E. Froustey, M. Unser, ”Deep Convolutional Neural Network for Inverse Problems in Imaging,” *IEEE Trans. Image Processing*, vol. 26, no. 9, pp. 4509-4522, 2017. **Best Paper Award**
- Neural networks: Deep spline framework
  - M. Unser, “A Representer Theorem for Deep Neural Networks,” *Journal of Machine Learning Research*, vol. 20, no. 110, pp. 1-30, 2019.
  - P. Bohra, J. Campos, H. Gupta, S. Aziznejad, M. Unser, “Learning Activation Functions in Deep (Spline) Neural Networks,” *IEEE Open Journal of Signal Processing*, Vol. 1 , pp. 295-309, 2020.
  - S. Ducotterd, A. Goujon, P. Bohra, D. Perdios, S. Neumayer, M. Unser “Improving Lipschitz-Constrained Neural Networks by Learning Activation Functions,” *Journal of Machine Learning Research*, vol. 25 (65), pp. 1–30, 2024.

ACKNOWLEDGMENTS

Many thanks to (former) members of EPFL's Biomedical Imaging Group

- Prof. Kyong Jin
- Dr. Shayan Aziznejad
- Dr. Thomas Debarre
- Stanilas Ducotterd
- Dr. Alexis Goujon
- Dr. Pakshal Bohra
- Prof. Sebastian Neumayer
- Dr. Mike McCann
- Dr. Dimitris Perdios
- Prof. Jaejun Yoo
- Prof. Matthieu Guerquin-Kern
- ….

CIBM
Center for Biomedical Imaging

and collaborators ...

- Prof. Demetri Psaltis
- Prof. Marco Stampanoni
- Prof. Carlos-Oscar Sorzano
- Prof Jianwei Ma
- ….

erc

FunLearn