Fast Training of Implicit Networks with Applications in Inverse Problems

Linghai Liu[1], Shuaicheng Tong[2], Lisa Zhao[3]

[1]Brown University
[2]University of California, Los Angeles
[3]University of California, Berkeley

EMORY

Introduction

Recent efforts in deep learning have turned towards solving inverse problems in imaging. For instance, Deep CNN was proposed for image denoising [12]. Moreover, the newly proposed implicit deep neural networks [2] are competitive with traditional feed-forward networks on sequential data [1] and are effective in inverse problems in imaging [4]. Implicit networks backpropagate through a fixed point, which allows them to maintain constant memory costs. However, they are expensive to train since backpropagating through implicit networks requires the computation of a Jacobian-based linear system for every gradient evaluation. Recently, a Jacobian-Free Backpropagation (JFB) approach was proposed to avoid solving the Jacobian-based system [3], which adopts an approximation of the true gradient.

Implicit Deep Learning

Given a dataset $\{(d_i, x_i)\}_{i=1}^N \subset \mathbb{R}^n \times \mathbb{R}^n$, the relation between the ground truths x_i's and our measurements d_i's is represented by the forward model [8]:

$$d_i = \mathcal{A}x_i + \varepsilon \tag{1}$$

where $\mathcal{A}$ is a (non)linear measurement operator and ε is random **unknown** noise.
Our goal is to design a weight-tying neural network $\mathcal{N}_\Theta : \mathbb{R}^n \mapsto \mathbb{R}^n$ with K layers, where each layer $T_\Theta : \mathbb{R}^n \mapsto \mathbb{R}^n$ is a (potentially nonlinear) mapping.
Given an input pair (d_i, x_i), we start with an initial guess x_i^0. Mimicking gradient descent and employing the forward model, we use the following updating rule [4]:

$$x_i^{k+1} = \underbrace{x_i^k - \eta\left(\nabla_x \|\mathcal{A}x_i^k - d_i\|_{L^2}^2 + S_\Theta(x_i^k)\right)}_{:=T_\Theta(x_i^k)} \tag{2}$$

where $\eta > 0$ is the step size and $S_\Theta : \mathbb{R}^n \mapsto \mathbb{R}^n$ is a trainable network that **learns** the gradient of an arbitrary regularizer. This is called the deep unrolling (DU) method.
For implicit networks, we expect the sequence $\{x_i^k\}_{k\in\mathbb{N}}$ to converge to a fix point x_i^* of T_Θ, i.e. $x_i^* = T_\Theta(x_i^*)$. This is true when T_Θ is a contraction mapping with Lipschitz constant $\gamma \in [0, 1)$.
Then we define

$$\mathcal{N}_\Theta(d_i) := x_i^* = T_\Theta(x_i^*) \tag{3}$$

as the output of our neural network, given an input d_i.
We can also choose other schemes to replace the iteration in Eq. 2, such as *proximal gradient descent* and *the alternating direction method of multipliers (ADMM)* [4].
Implicit neural networks can be trained using gradient descent and a calculated fix point. Suppose an experimenter chooses loss function ℓ. Then using implicit differentiation and Eq. 3 we have:

$$\frac{d\ell}{d\Theta} = \frac{d\ell}{d\mathcal{N}_\Theta}\frac{d\mathcal{N}_\Theta}{d\Theta} = \frac{d\ell}{d\mathcal{N}_\Theta}\frac{dx^*}{d\Theta} = \frac{d\ell}{d\mathcal{N}_\Theta}\left(I - \frac{dT_\Theta(x^*;d)}{dx^*}\right)^{-1}\frac{\partial T_\Theta(x^*;d)}{\partial\Theta} \tag{4}$$

Eq. 4 calculates the true gradient of our neural network parameters Θ with respect to loss function ℓ. However, calculating the inverse

$$\left(I - \frac{dT_\Theta(x^*)}{dx^*}\right)^{-1}$$

is **highly nontrivial** since a Jacobian-based linear system needs to be solved.

Jacobian-Free Backpropagation (JFB)

The goal of JFB is to **alleviate memory requirement** and **avoid high computational cost** in implicit networks. The key idea is to replace the problematic Jacobian $\left(I - \frac{dT_\Theta(x^*)}{dx^*}\right)$ in Eq. 4 with the identity matrix I. As a result, implicit networks are trained faster and more easily implemented—all while maintaining competitive results in image classification tasks [3].
We make the proposed substitution in Eq. 4 to approximate the gradient $\frac{d\ell}{d\Theta}$ and obtain:

$$p_\Theta = \frac{d\ell}{d\mathcal{N}_\Theta}\frac{\partial T_\Theta(x^*)}{\partial\Theta}$$

which is a descent direction for the loss ℓ.
Note: the JFB approach relies on more assumptions to hold:

- T_Θ is continuously differentiable w.r.t. Θ
- $M := \frac{\partial T_\Theta}{\partial\Theta}$ has full column rank.
- M is well-conditioned, i.e., $\kappa(M^TM) < \frac{1}{\gamma}$, where γ is the Lipschitz constant of T_Θ.

Results

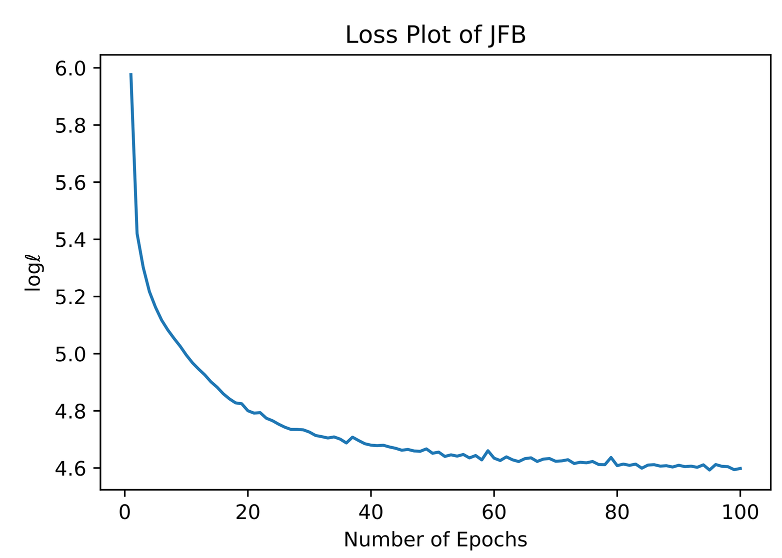

Fig. 1: Plot of Mean Squared Error (MSE) Per Image
step size $\eta = 10^{-3}$, learning rate $\alpha = 10^{-4}$

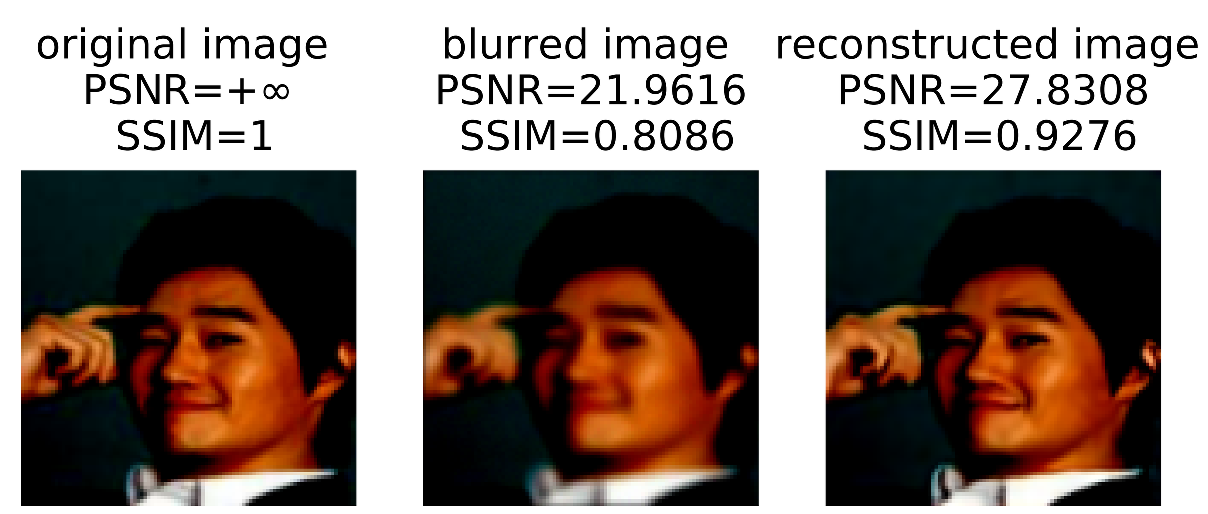

Fig. 2: Result of Proposed JFB on a Test Image Used by [4]

Note: Two metrics are commonly used for assessing the quality of reconstructed images [5]: the peak-signal-to-noise ratio (PSNR, a positive number, best at $+\infty$) and the structural similarity index measure (SSIM, also positive, best at 1).

Comparison

	Total Variation [9]	Plug-n-Play [10]	Deep Equilibrium [4]	JFB (Ours)
PSNR	26.79	29.77	32.43	27.83
SSIM	0.86	0.88	0.94	0.9276

The table above records the mean PSNR and SSIM of the dataset for our various models (statistics from [4]). It can be observed that applying JFB to training models for inverse problems in imaging is competitive.

Remarks

Our model is currently trained on a subset (8,000 images) of the CelebA dataset [7] using 1 NVIDIA RTX A6000 GPU.
Future directions include: (i) continuing to train current model until convergence (ii) training JFB models on other schemes (proximal gradient descent & ADMM) as in [4] (iii) training JFB models on datasets such as fastMRI [6] [11]

Acknowledgements

We sincerely thank the guidance of our mentor, Dr. Samy Wu Fung, and other mentors at Emory University for the opportunity. This work is supported in part by the US National Science Foundation awards DMS-2051019 and DMS-1751636.

References

[1] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. "Deep equilibrium models". In: *Advances in Neural Information Processing Systems* 32 (2019).

[2] Laurent El Ghaoui et al. "Implicit deep learning". In: *SIAM Journal on Mathematics of Data Science* 3.3 (2021), pp. 930–958.

[3] Samy Wu Fung et al. "JFB: Jacobian-Free Backpropagation for Implicit Networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.6 (June 2022), pp. 6648–6656.

[4] Davis Gilton, Gregory Ongie, and Rebecca Willett. "Deep equilibrium architectures for inverse problems in imaging". In: *IEEE Transactions on Computational Imaging* 7 (2021), pp. 1123–1133.

[5] Alain Hore and Djemel Ziou. "Image quality metrics: PSNR vs. SSIM". In: *2010 20th international conference on pattern recognition*. IEEE. 2010, pp. 2366–2369.

[6] Florian Knoll et al. "fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning". In: *Radiology: Artificial Intelligence* 2.1 (2020), e190007.

[7] Ziwei Liu et al. "Deep learning face attributes in the wild". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 3730–3738.

[8] Gregory Ongie et al. "Deep learning techniques for inverse problems in imaging". In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 39–56.

[9] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. "Nonlinear total variation based noise removal algorithms". In: *Physica D: Nonlinear Phenomena* 60.1 (1992), pp. 259–268.

[10] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. "Plug-and-play priors for model based reconstruction". In: *2013 IEEE Global Conference on Signal and Information Processing*. IEEE. 2013, pp. 945–948.

[11] Jure Zbontar et al. "fastMRI: An open dataset and benchmarks for accelerated MRI". In: *arXiv preprint arXiv:1811.08839* (2018).

[12] Kai Zhang et al. "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising". In: *IEEE transactions on image processing* 26.7 (2017), pp. 3142–3155.