Convolutional Neural Networks for Direct Text Deblurring

Michal Hradiš[1]
ihradis@fit.vutbr.cz
Jan Kotera[2]
kotera@utia.cas.cz
Pavel Zemčík[1]
zemcik@fit.vutbr.cz
Filip Šroubek[2]
sroubekf@utia.cas.cz

[1] Faculty of Information Technology
Brno University of Technology
Brno, Czech Republic

[2] Institute of Information Theory and Automation
Czech Academy of Sciences
Prague, Czech Republic

We investigate if convolutional neural networks (CNN) can learn to directly perform blind image deconvolution and restoration – that is if they can provide high-quality restored images directly from blurred inputs without any knowledge of the specific degradataion process. In our experiments on text documents, CNNs significantly outperformed existing blind deconvolution methods, including those optimized for text, in terms of image quality and OCR accuracy. In fact, the convolutional networks outperformed even state-of-the-art non-blind methods for anything but the lowest noise levels.

The architectures we use are inspired by the very successful networks that recently redefined state-of-the-art in many computer vision tasks starting with image classification on ImageNet by Krizhevsky et al. [2]. The networks are composed of multiple layers of convolutions and element-wise Rectified Liear Units (ReLU):

$$\begin{aligned} F_0(y) &= y \\ F_l(y) &= \max(0, W_l * F_{l-1}(y) + b_l), l = 1, \ldots, L-1 \\ F(y) &= W_L * F_{L-1}(y) + b_L \end{aligned} \tag{1}$$

The input and output are both 3-channel RGB images with values mapped to interval $[-0.5, 0.5]$. Each layer applies c_l convolutions with filters spanning all channels c_{l-1} of the previous layer. The last layer is linear (without ReLU).

As in previous works [5], we train the networks by minimizing mean squared error on a dataset $D = (x_i, y_i)$ of corresponding clean and corrupted image patches:

$$\arg\min_{W,b} \frac{1}{2|D|} \sum_{(x_i,y_i) \in D} ||F(y_i) - x_i||_2^2 + 0.0005||W||_2^2 \tag{2}$$

We evaluated the approach on a large set of documents from the CiteSeerX repository which we rendered at 120-150 DPI. A dataset was created by aplying random geometric transformations simulating deviations from optimal camera position to small rendered page regions and by convolving with realistic de-focus and camera-shake blur kernels (distribution of kernel sizes is show in Figure 1 right-bottom). We purposely limited the image degradations to shift-invariant blur and additive noise to allow for fair comparison with the baseline methods, which are not designed to handle other aspects of an imaging process.

In our experiments, deeper networks performed singnificantly better and restoration quality was fairly insensitive to other architectural choices. This can be clearly seen in Figure 1. Top-left shows best the benefit of deeper networks. Top-right shows that performance is insensitive to the size of the first layer filters. Bottom-left shows that networks with more channels per layer perform slightely better, but the benefit is negligable compared to increased network depth.

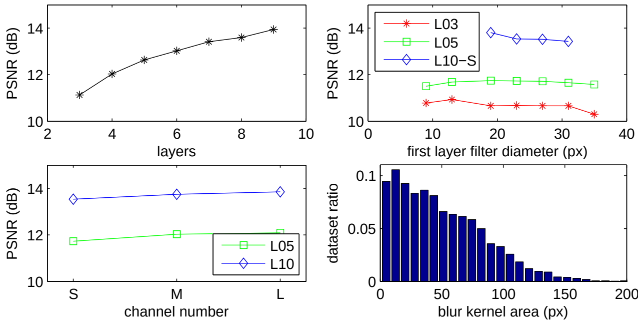

Figure 1: Different CNN architectures. The number after L (e.g. L10) indicates number of layers in the network. top-left – network depth; top-right – spatial support size; bottom-left – channel number; bottom-right – distribution of blur-kernel sizes in dataset

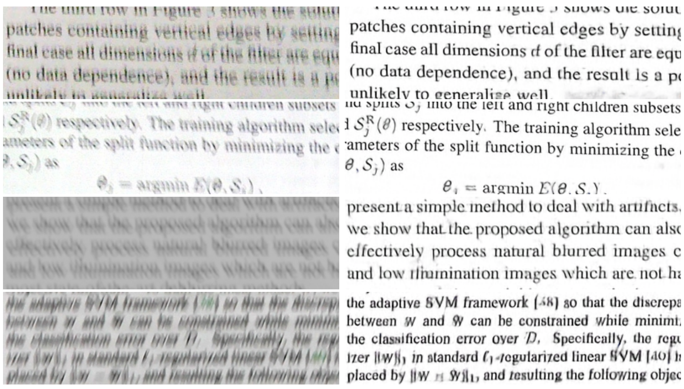
patches containing vertical edges by setting
final case all dimensions d of the filter are equ
(no data dependence), and the result is a p
unlikely to generalise well.
nd splits S_j into the left and right children subsets
d S_j^R(θ) respectively. The training algorithm sele
ameters of the split function by minimizing the
θ, S_j) as
present a simple method to deal with artifacts.
we show that the proposed algorithm can also
effectively process natural blurred images c
and low illumination images which are not h
the adaptive SVM framework [48] so that the discrepa
between w and w can be constrained while minimi
the classification error over D. Specifically, the regu
izer ||w||_1 in standard ℓ_1-regularized linear SVM [40] i
placed by ||w − w||_1, and resulting the following objec

Figure 2: CNN deblurring results of challenging real images.

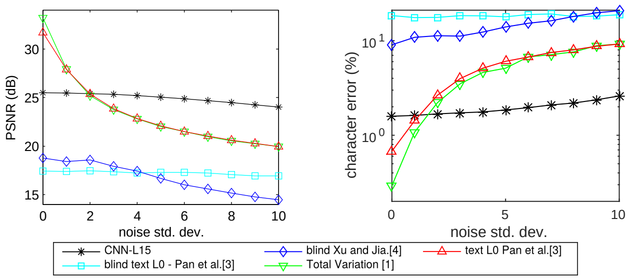

Figure 3: Blind deconvolution image quality (left) and OCR error on reconstructed images (right).

The largest and deepest network we trained has 15 layers with 2.3M parameters (PSNR 16.06 dB on the dataset from Figure 1). We compared this network with two blind [3, 4] and two non-blind [1, 3] methods on 200×200 image patches extracted from unseen documents (see Figure 3). The methods of Pan et al. [3] were designed specifically for text images. The CNN clearly outperforms the blind methods for all noise levels and the non-blind methods for all but the lowest noise levels. Surprisingly, the CNN maintains good quality even for noise levels higher than for what it was trained for.

[1] Stanley H. Chan, Ramsin Khoshabeh, Kristofor B. Gibson, Philip E. Gill, and Truong Q. Nguyen. An augmented Lagrangian method for total variation video restoration. *IEEE TIP*, 20(11):3097–3111, 2011.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*, 2012.

[3] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring Text Images via L0-Regularized Intensity and Gradient Prior. In *CVPR*, 2014.

[4] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *ECCV*, 2010.

[5] Li Xu, Jimmy SJ. Ren, Ce Liu, and Jiaya Jia. Deep Convolutional Neural Network for Image Deconvolution. In *NIPS*, 2014.