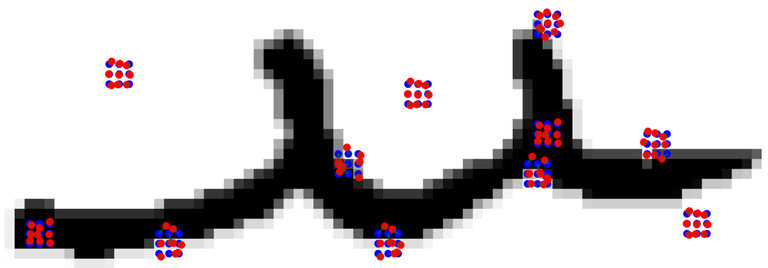
Watch Your Strokes: Improving Handwritten Text Recognition with Deformable Convolutions

Iulian Cojocaru, Silvia Cascianelli, Lorenzo Baraldi,
Massimiliano Corsini, Rita Cucchiara
University of Modena and Reggio Emilia
Email: {name.surname}@unimore.it

Overview

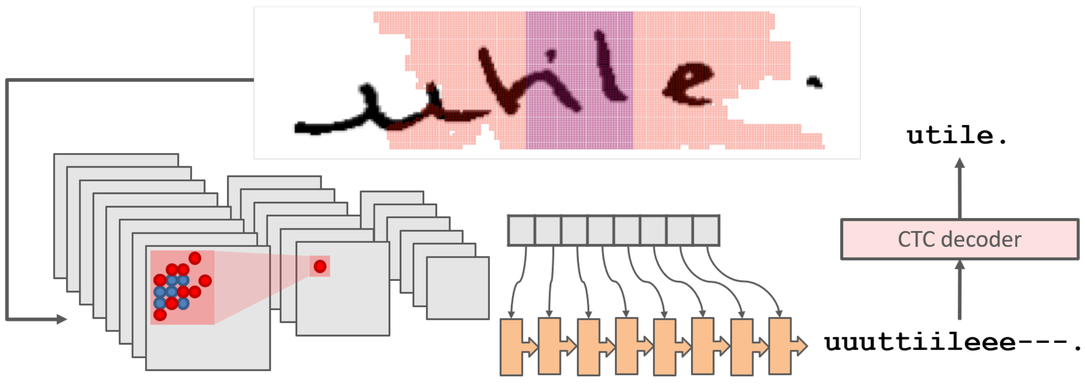
State-of-the-art approaches for **Handwritten Text Recognition** (HTR) in free-layout pages usually encode input images with Convolutional Neural Networks, whose kernels are typically defined on a fixed grid and focus on all input pixels independently. However, **handwritten texts are a sparse structure**, in which only a small part of the input (*i.e.* the ink pixels) is useful for recognition. Moreover, handwritten characters and words vary in shape, scale, and orientation but, with standard convolutions, this variability is not effectively taken into account unless ad hoc data augmentation or preprocessing is performed.



→ We propose to apply **deformable convolutions** (DefConvs) [2] in place of standard convolutions for the HTR task.

Full-DefConv HTR Model

We adapt the sequence recognition network proposed in [5], commonly used as a base for HTR schemes, and replace all its standard convolution layers with DefConv layers. The model consists of three main parts: a CNN to extract sequences of features from the input image, an RNN to produce labels' probabilities based on the sequence, and a decoding block to output the final transcription. The network is trained to maximize the Connectionist Temporal Classifier (CTC) probability of the transcribed sequence.



Formally, given a kernel k of learnable weights and a regular grid $\mathcal{N}$, the DefConv on a pixel p is:

$$y(p) = \sum_{d \in \mathcal{N}} k(d) \cdot x(p + d + \Delta d), \quad (1)$$

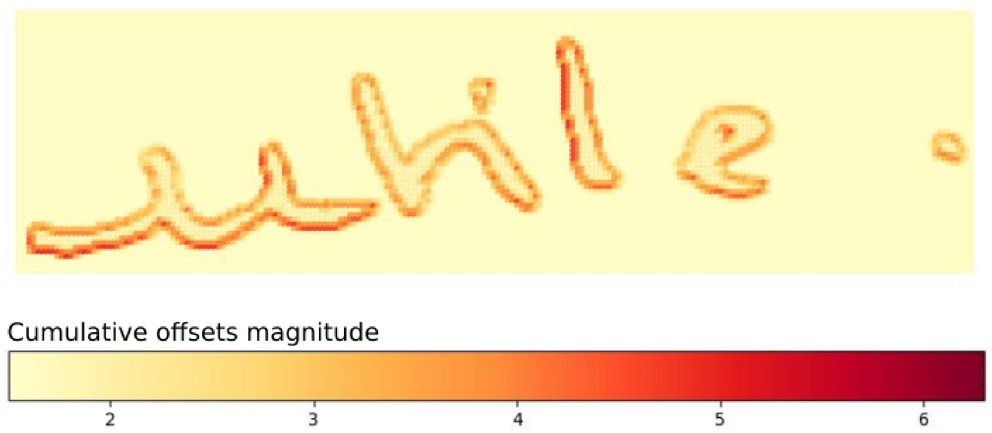
where d is a displacement vector, and Δd is a 2D offsets vector. The offsets are learned alongside the kernel weights in an additional convolutional layer, thus ensuring a **content-dependent deformation**.

References

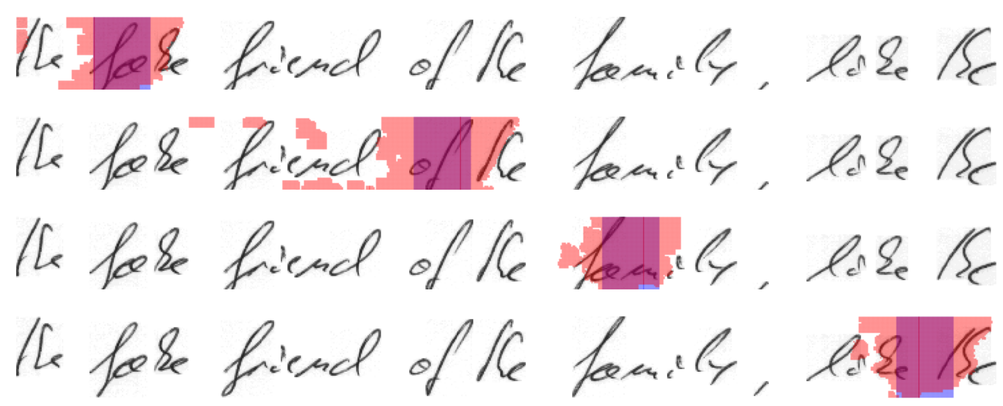
[1] T. Bluche. Joint line segmentation and transcription for end-to-end handwritten paragraph recognition. In *NeurIPS*, 2016.
[2] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *CVPR*, 2017.
[3] V. Pham, T. Bluche, C. Kermorvant, and J. Louradour. Dropout improves recurrent neural networks for handwriting recognition. In *ICFHR*, 2014.
[4] J. Puigcerver. Are multidimensional recurrent layers really necessary for handwritten text recognition? In *ICDAR*, 2017.
[5] B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. PAMI*, 39(11):2298–2304, 2016.
[6] P. Voigtlaender, P. Doetsch, and H. Ney. Handwriting recognition with large multidimensional long short-term memory recurrent neural networks. In *ICFHR*, 2016.
[7] C. Wigington, C. Tensmeyer, B. Davis, W. Barrett, B. Price, and S. Cohen. Start, follow, read: End-to-end full-page handwriting recognition. In *ECCV*, 2018.

DefConvs on Handwtitten Text Images

DefConvs kernel deforms to focus on the writing instead of the background. Indeed, the kernel's grids sampling in uniform regions are less deformed than those sampling on the edges of the writing parts.



Compared to an HTR network using standard convolutions, the receptive fields of our Full-DefConv model are non-connected areas of irregular shape that better **adapt to handwritten strokes** and cover a wider portion of the image thanks to the limited amount of additional offsets parameters.



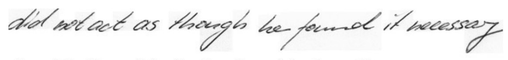
Results

Compared to the baseline (Shi *et al.*[5]), Full-DefConv allows **decreasing both the CER and the WER**. Full-DefConv also performs competitively w.r.t. other State-of-the-Art approaches, especially those that, as in our case, do not perform any preprocessing or data augmentation.

From a qualitative viewpoint, **background irregularities do not affect the transcription** produced by Full-DefConv. This is also confirmed by the quantitative comparison between our approach and the baseline tested on images with added White Gaussian noise or Poisson shot noise with different variance.

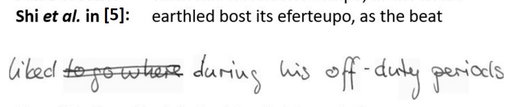
Method	IAM dataset CER	IAM dataset WER	RIMES dataset CER	RIMES dataset WER
Full-DefConv	4.6	19.3	4.6	14.8
Shi *et al.*[5]	5.7	23.2	5.3	17.5
Wigington *et al.*[7]	6.4	23.2	2.1	9.3
Voigtlaender *et al.*[6] – LM	8.3	27.5	4.0	17.7
Puigcerver [4]	6.2	20.2	2.6	10.7
Bluche [1]	7.9	24.6	2.9	12.6
Pham *et al.*[3]	10.8	35.1	6.8	28.5

IAM dataset:



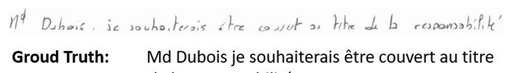
Groud Truth: did not act as though he found it necessary
Full-DefConv: did not act as though he found it necessay
Shi *et al.* [5]: dd n act as thaugh kefanod it necarseay

Groud Truth: earth had lost its life-tempo, as the heart
Full-DefConv: earth had lost its lefe-tempo, as the heart
Shi *et al.* in [5]: earthled bost its eferteupo, as the beat

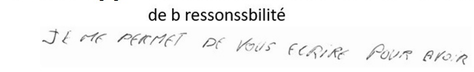


Groud Truth: liked during his off-duty periods
Full-DefConv: liked # during his off-duty periods
Shi *et al.* [5]: liked tegotere during his off-duty periots

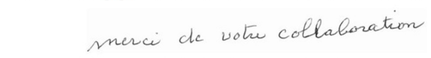
RIMES dataset:



Groud Truth: Md Dubois je souhaiterais être couvert au titre de la responsabilité
Full-DefConv: Ma Duois je souhaiterais être coupet au. Titre de l ressonsabilité
Shi *et al.* in [5]: ma Bubois. Je souhatersir être lea mert u titre de b ressonssbilité



Groud Truth: JE ME PERMETS DE VOUS ECRIRE POUR AVOIR
Full-DefConv: JE ME MERMETS DE VOUS ECPIR POUP QNOIR
Shi *et al.* in [5]: JEe PERMET DE voUS FcAIrRe Pour avoin



Groud Truth: merci de votre collaboration
Full-DefConv: erci de votre collaloration
Shi *et al.* in [5]: A'acexio de votre collaloration

Noise		IAM dataset Full-DefConv CER	IAM dataset Full-DefConv WER	IAM dataset Shi *et al.*[5] CER	IAM dataset Shi *et al.*[5] WER	RIMES dataset Full-DefConv CER	RIMES dataset Full-DefConv WER	RIMES dataset Shi *et al.*[5] CER	RIMES dataset Shi *et al.*[5] WER
$\mathcal{G}(0, 10)$		4.7	19.5	5.8	23.7	4.6	14.8	5.3	17.3
$\mathcal{G}(0, 20)$		5.5	22.2	6.9	26.5	4.7	15.4	5.4	18.2
$\mathcal{G}(0, 30)$		18.3	49.0	24.4	62.8	5.1	17.0	6.0	20.2
$\mathcal{P}(0, 10)$		4.8	19.8	5.9	24.0	4.6	14.8	5.3	17.4
$\mathcal{P}(0, 20)$		5.5	22.0	6.7	26.0	4.6	15.1	5.4	17.7
$\mathcal{P}(0, 30)$		10.6	33.3	13.6	41.2	4.7	15.1	5.5	18.2