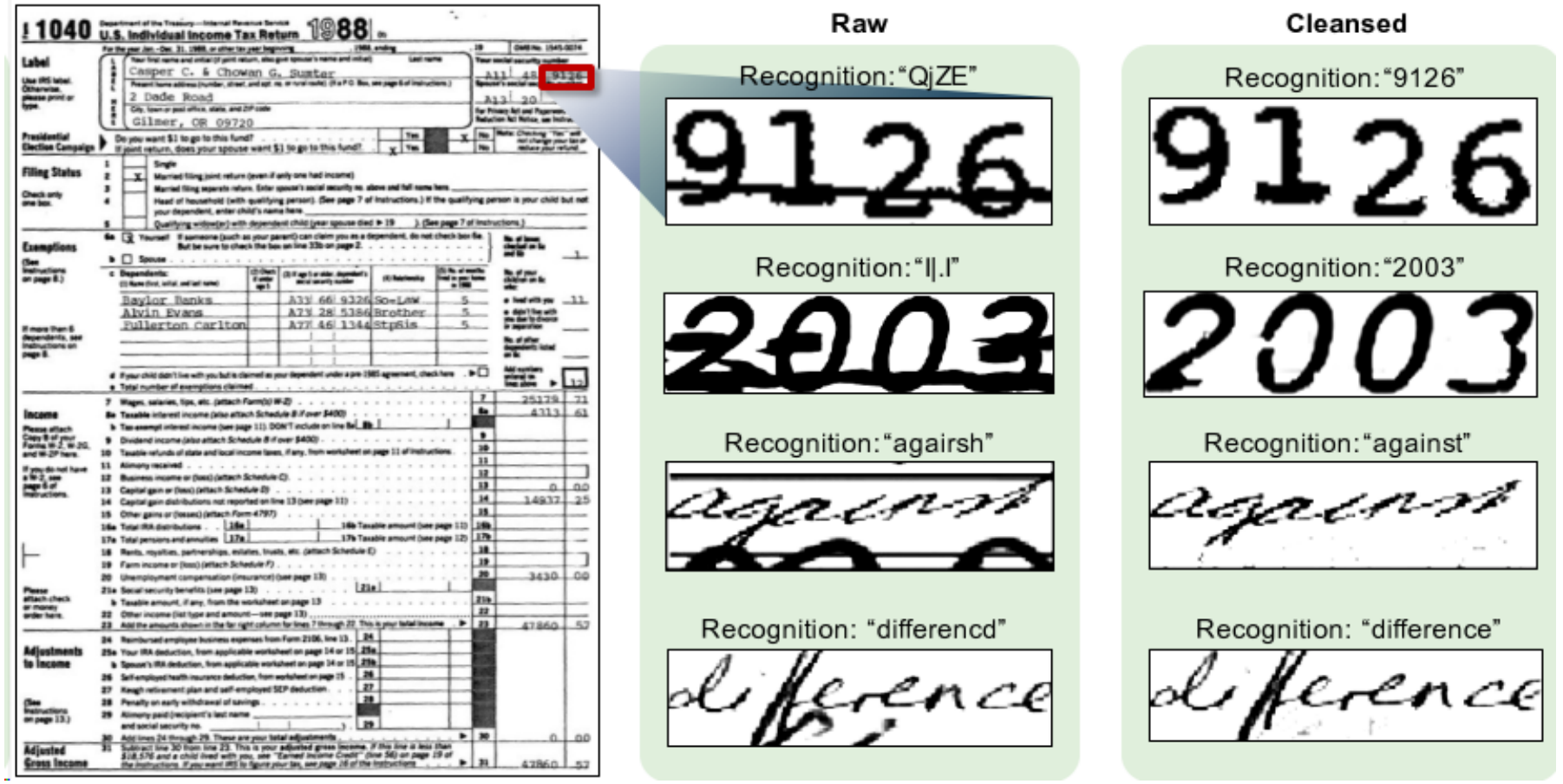
Document Intelligence: Deeplearning Artifacts Removal Model

NeurIPS 2019 Paper

Oct, 2019

DeepErase - Deep Learning for Artifacts Removal
Background

Paper-intensive industries like insurance and law have long leveraged optical character recognition (OCR) to transcribe scanned documents into text strings for downstream processing. Text extracted from real world documents is often nested with tabular structures of underlines, boxes or even strikes of ink from neighbor form cells. Such ink artifacts can severely interfere with the performance of recognition algorithms or other downstream processing tasks.



DeepErase - Deep Learning for Artifacts Removal
Model Framework

We devised a synthetic training data generation method and trained a deeplearning-based segmentation model to remove the artifacts in form crops. The practice enhances the OCR performance. The segmentation accuracies for both frameworks were above 96% based on 280,000 labeled images. The cleaning decreased tesseract recognition character error rate from 97.26% to 59.95% based on NIST sd02 Tax Form dataset crops. The model paper has been adopted in NeurIPS 2019*.

1 Input Crops with Artifacts → **2 Deep Learning Model (SegNet or UNet)** → **3 Output and Assessment**

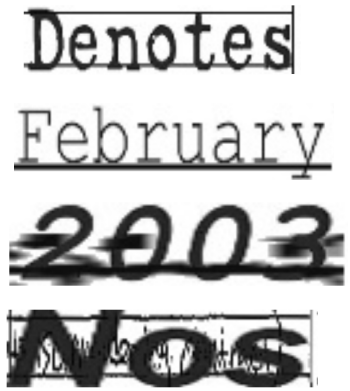
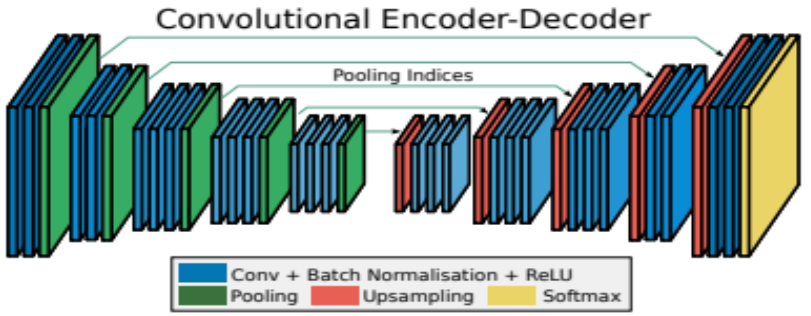
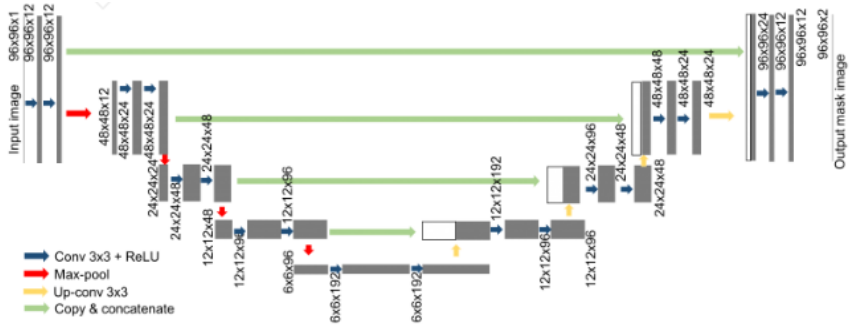


Image Resize Normalization

Model 1: SegNet



Model 2: UNet



Model Outputs

Result Validation

- Overall segmentation accuracy 96%
- Artifacts pixel capture rate 98%
- Non-artifacts pixel capture rate 96%

DeepErase: Weakly Supervised Ink Artifact Removal in Document Text Images, Yike Qi, W Roony Huang*, Qianqian Li, Jonathan Degange

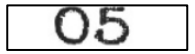
Deep Learning for Artifacts Removal
Synthetic Training Data Generation & Usage

In order to automatically obtain a corpus of dirty images, we created a program to impose realistic-looking artifacts on the readily available datasets of clean images. We focus on four types of artifacts:

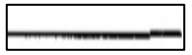
- machine-printed underlines
- Machine-printed fill-in-the-blank boxes
- Random smudges
- Handwritten spurious strokes

Below is a example of a base image and an artifact used in the assembly process.

Base Image:



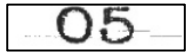
Artifact Line:



Patched Image:

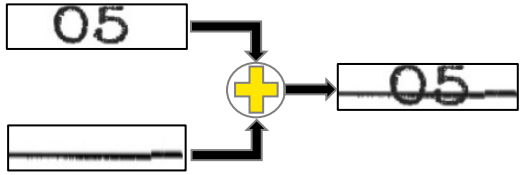
Predicted Mask:

Artifact Removed:

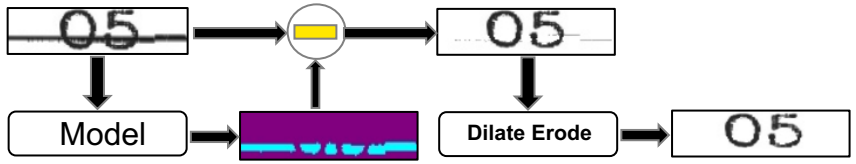


Slight Dilate Erode:

Artifact Patched Training Data Generation



Artifact Removal Pipeline



Deep Learning for Artifacts Removal
Apply to OCR

A typical OCR process at least includes two major modules: Text Detection + Text Recognition. Between the two modules, text segmentation and quality enhancement process may be applied. Below is a demonstration of inhouse OCR design. The model is inserted as a preprocessing layer between 2 and 3.

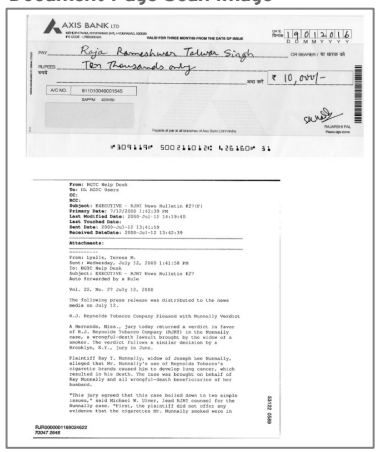
1 **Input Image Prepocess**

2 **Text Detection (EAST)**

3 **Text Recognition (DenseNet CNN + LSTM + CTC)**

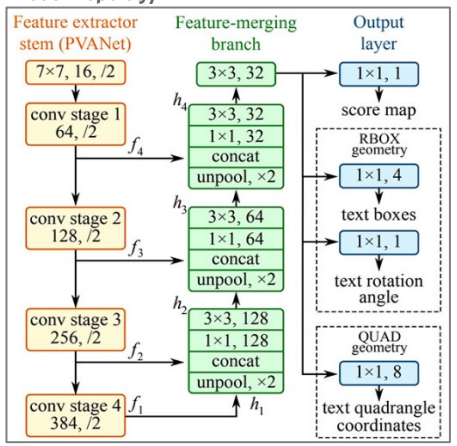
4 **Paragraph Recovery**

Document Page Scan Image

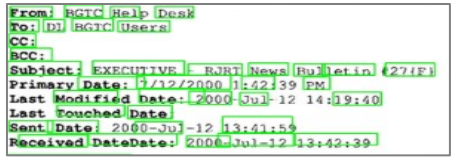


Document Resize Normalization

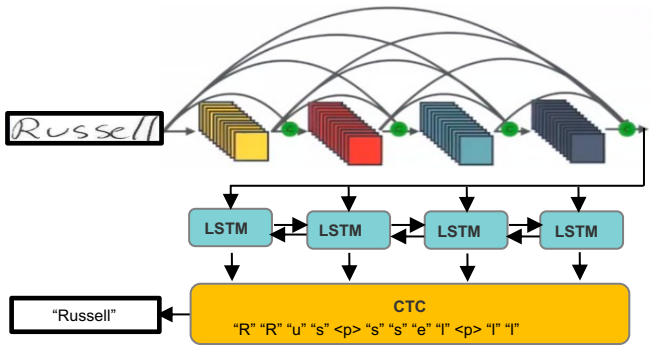
Model Topology



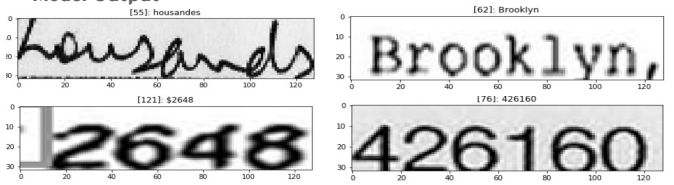
Model Output



Model Topology



Model Output



Paragraph Construction

- Document text layouts are recovered based on word bounding boxes coordinates and recognized word values

Assessment and Downstream Process

Downstream:

- Document Classification
- Information Extraction
- Customer Application Assessment

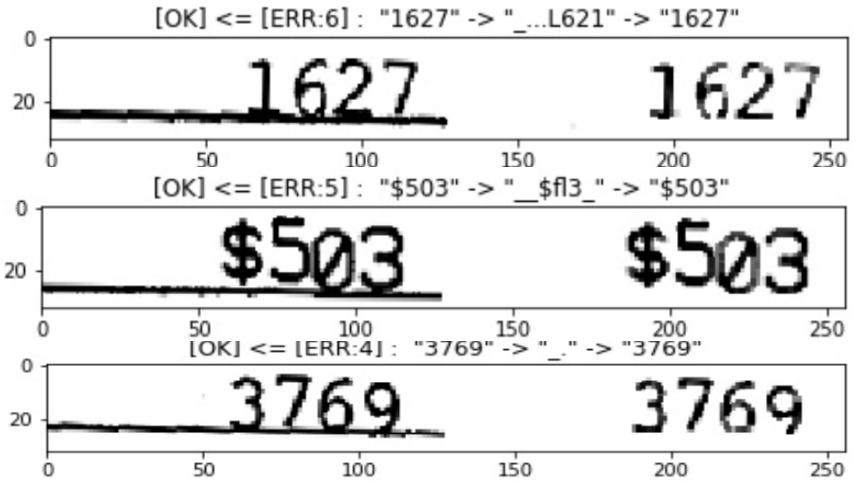
DeepErase - Deep Learning for Artifacts Removal
Snapshot of Model Performance

Performance:

- Able to identify line/box artifacts with high accuracy
- **Unet 96.7%** of over all pixel level accuracy
- **Unet 96.6%** of non-artifact pixel capture rate
- **Unet 98.5%** of artficat pixel capture rate

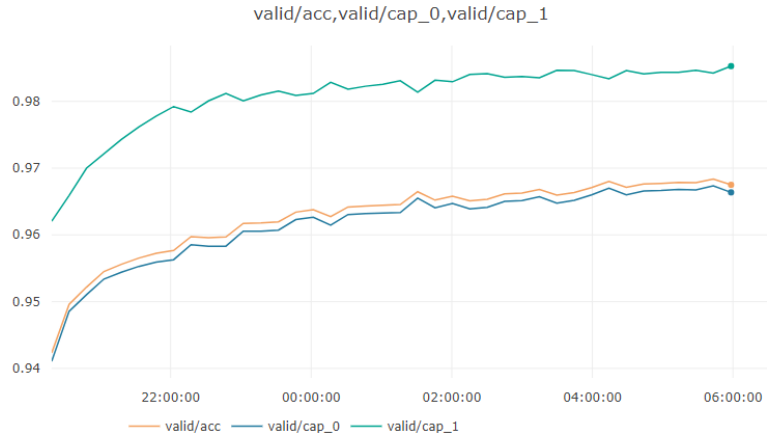
- Based on a public out of distribution IRS extraction datasets, the model reduces tesseract recognition character error rate from **97.26%** to **59.95%**

Sample Predicted Output:

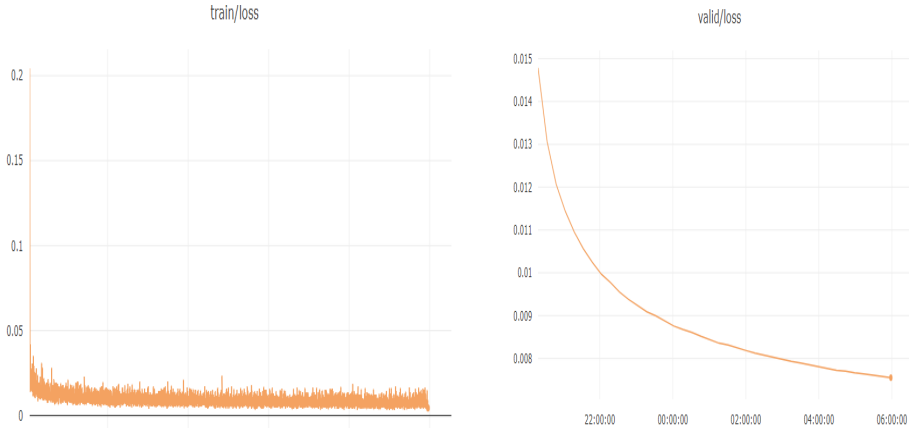


* [after cleaning] <=[before cleaning]: “ground truth”->”before cleaning” ->”after cleaning”

Model Performance:



Fast Convergence of Loss with Training Iteration:



DeepErase - Deep Learning for Artifacts Removal
Additional Development and Assessment

Additional Development Work:
- Joint Training: DeepErase + Recognition

Benchmarking:
- Houghline Method*
- Manual Supervision Method**

Extensive Recognition Testing:
- Printed Validation Dataset
- Handwritten Validation Dataset
- Printed IRS Crop Images (NIST sd02)
- Handwritten Crop Images (NIST sd06)

* L. Likforman-Sulem, A. Hanimyan, and C. Faure. A hough based algorithm for extracting text lines in handwritten documents. In Proceedings of 3rd International Conference on Document Analysis and Recognition, volume 2, pages 774–777. IEEE, 1995. 2

** J. Calvo-Zaragoza, G. Vigliensoni, and I. Fujinaga. One-step detection of background, staff lines, and symbols in medieval music manuscripts with convolutional neural networks. In ISMIR, pages 724–730, 2017.

Thanks so much