Document analysis and recognition: The Deep Learning Era

Yousri Kessentini
Digital Research center of Sfax

yousri.kessentini@crns.rnrt.tn
http://yousri.fr.cr
About me

- Yousri Kessentini
  - Assistant professor at CRNS
  - Head of DeepVision research team
  - More than 15 years experience in the document analysis & recognition
    - PHD in the university of Rouen
    - Post-doc Itesoft-LITIS
    - Researcher in CRNS

Certified as an official instructor & ambassador from NVIDIA Deep Learning Institute

HomePage: http://yousri.fr.cr
Digital research center of Sfax

- **11 000 m² / 165 researchers**

**Research Topics**

1. Computer and distributed systems
2. Image and signal processing
3. Document and data analysis
4. Recognition of shapes and objects (3D)
5. Electronics and embedded systems
6. Computer Networks
7. Security of computer systems

**Application Domains**

- Digital Health
- Smart Agriculture
- Industry 4.0
- Smart Grid
Research activities

DeepVision

**Fields**: Pattern recognition, Computer vision, Machine learning

- **Image processing**
  - Object detection
  - Image recognition
  - Face recognition

- **Document processing**
  - Document Analysis
  - Handwriting recognition
  - OCR, word-spotting
Introduction

A huge amount of documents (machine printed, handwritten)

• Preservation
• Storage
• Access to contents

➔ High Cost of manual processing

PayStream Advisors research reveals the average cost of manually processing an invoice can be as high as $20, versus $4 for automated invoice
Introduction

- Automatic document processing
  - Classification
  - Document Enhancement
  - Writer/script Identification
  - Layout analysis
  - Recognition
  - Spotting ...

- Difficulties
  - Unstructured data
  - Deformations, noise
  - Different handwriting styles
  - Segmentation problem
Document Recognition process

Original document → Document enhancement → Line segmentation

Script identification → Document recognition

Preprocessing
Feature Extraction
Optical Model
Langage Model

Deep Learning

NN-HMM

HMM
Handwriting recognition before deep Learning

GMM-HMM with carefully chosen features

Combination of BLSTMs at different levels


30/10/2020
Deep Learning arrived

- Applied directly to the pixel of the raw text line image
  - BLSTM → MDLSTM → CRNN → Seq2Seq ...

CRNN : CNN+BLSTM

MDLSTM

Seq2Seq model
HIGH PERFORMANCE WITH DL
BUT…

How to recognize degraded documents?
Document Enhancement

- Use Generative Adversarial Networks (GAN) to restore severely degraded document images.

Components of GANs

The discriminator tries to identify real data from fakes created by the generator.

The generator turns noise into an imitation of the data to try to trick the discriminator.

Train GAN jointly via minimax game:

- The discriminator tries to maximize its classification accuracy.
- The generator tries to minimize the discriminator’s classification accuracy.

\[ \min_G \max_D V(D, G) \quad V(D, G) = \mathbb{E}_{x \sim p(x)}[\log D(x)] + \mathbb{E}_{z \sim q(z)}[\log(1 - D(G(z)))] \]

Maximized by D  Minimized by G
DE-GAN: Document Enhancement GAN

Generator

Discriminator

Degraded images

Enhanced images

Ground-truth images

Log loss

Discriminator

Real / Fake

30/10/2020
DE-GAN: Document Enhancement

Original

Ravno za to nam je pa
Samo, da bi se mož zato p
težavno vredništvo na svoje r
iskrenih domoljubov, ki pozn
stva, mu se gotovo manjkalo

Niblack [50]

Ground truth

Ravno za to nam je pa
Samo, da bi se mož zato p
težavno vredništvo na svoje r
iskrenih domoljubov, ki pozn
stva, mu se gotovo manjkalo

Sauvola et al. [51]

Otsu [16]

Ravno za to nam je pa
Samo, da bi se mož zato p
težavno vredništvo na svoje r
iskrenih domoljubov, ki pozn
stva, mu se gotovo manjkalo

DE-GAN
### DE-GAN: Watermark Removal

<table>
<thead>
<tr>
<th>Watermarked Images</th>
<th>Ground Truth</th>
<th>Predicted Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept of evolution as presented in the book</td>
<td>Concept of evolution as presented in the book</td>
<td>Concept of evolution as presented in the book</td>
</tr>
<tr>
<td>By the time it was discovered, the concept was well-established.</td>
<td>By the time it was discovered, the concept was well-established.</td>
<td>By the time it was discovered, the concept was well-established.</td>
</tr>
<tr>
<td>It was the work of Charles Darwin that led to the discovery of evolution theory.</td>
<td>It was the work of Charles Darwin that led to the discovery of evolution theory.</td>
<td>It was the work of Charles Darwin that led to the discovery of evolution theory.</td>
</tr>
<tr>
<td>Watermarked text:</td>
<td>Watermarked text:</td>
<td>Watermarked text:</td>
</tr>
<tr>
<td>Thus, for example, a carnivore consuming flesh. At the same time,</td>
<td>Thus, for example, a carnivore consuming flesh. At the same time,</td>
<td>Thus, for example, a carnivore consuming flesh. At the same time,</td>
</tr>
<tr>
<td>devouring prey; tearing and dividing its flesh; the entire</td>
<td>devouring prey; tearing and dividing its flesh; the entire</td>
<td>devouring prey; tearing and dividing its flesh; the entire</td>
</tr>
</tbody>
</table>

---

**Note:** The WATERMARKED text is redacted before redaction.
DE-GAN: Deblurring

### Source code: https://github.com/dali92002/DE-GAN
HIGH PERFORMANCE WITH DL

BUT…

Recognition still depends on linguistic resources

How to handle OOV words?
Out-of-vocabulary words

- Handwriting recognition systems usually rely on static dictionaries
  - Small lexicon: Low coverage rate
  - Big lexicon: High computing time + high confusion between words
- Full coverage of these dictionaries is generally not achieved

The word «الجامعات» Does not exist in the vocabulary

The most similar word from the dictionary

How to recognize OOV words?
Solution 1 : Combining different sub-word modeling

Solution 2: OOV words detection and recovery using dynamic dictionaries

**Idea:**
- Keep the IV words recognized by the word lexicon driven recognition
- Recover OOV word using dynamic lexicons built from large text corpora

<table>
<thead>
<tr>
<th>Line Image</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>نوفل</strong> بدأ <strong>موقف</strong> الحجيج الحاج آخر يلبى <strong>تلبية</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word lexicon driven recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>تلبية</strong> آخر حاج الحجيج نوفل بدأ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PAW lexicon driven recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>تلبية</strong> آخر حاج الحجيج نوفل بدأ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Morphemes lexicon driven recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>تلبية</strong> آخر حاج الحجيج نوفل بدأ</td>
</tr>
</tbody>
</table>

**OOV Detection stage**
Solution 2: OOV words detection and recovery using dynamic dictionaries

OOV words Recovery stage

1. Identification of the Reference OOV
2. Lexicon Extension
3. Dynamique Lexicon driven recognition
4. Line text hypothesis

Word lexicon driven recognition

- NOV

PAW lexicon driven recognition

- NOV

Morphemes Word lexicon driven recognition

- NOV

Results

- KHATT dataset
- Train : 9475 Test : 2007
- Lexicon : 18933 words, 11,46% OOV

<table>
<thead>
<tr>
<th>Systems</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamdani et al. [2]</td>
<td>26.80</td>
</tr>
<tr>
<td>BenZeghiba et al. [1]</td>
<td>30.9</td>
</tr>
<tr>
<td>BenZeghiba [3]</td>
<td>34.3</td>
</tr>
<tr>
<td>Our approach</td>
<td>20.83</td>
</tr>
</tbody>
</table>
HIGH PERFORMANCE WITH DL
BUT…

What about cases where you don't have labeled data?
Recognition of ciphered manuscript

- Around 1% of documents in archives contain encrypted text:
  - Diplomatic correspondence, secret societies/religious groups...

- DECODE/DECRYPT project:
  - To develop resources & tools for automatic decryption of enciphered documents from early modern times.

Problem: Very few (or none) labeled data to train
Few-shot Learning for Historical Ciphered Manuscript Recognition

Symbol detection step

Few-shot Learning for Historical Ciphered Manuscript Recognition

Text recognition step
Few-shot Learning for Historical Ciphered Manuscript Recognition

![Image showing train and test data]

<table>
<thead>
<tr>
<th>Method</th>
<th>Train (synthetic): Omniglot</th>
<th>Test (ciphers):</th>
<th>Copiale</th>
<th>Borg</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTMs</td>
<td>25</td>
<td>-</td>
<td>0.11</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>-</td>
<td>0.10</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>-</td>
<td>0.07</td>
<td>0.45</td>
</tr>
<tr>
<td>Few-shot</td>
<td>2</td>
<td>1</td>
<td>0.10</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0.10</td>
<td>0.18</td>
</tr>
</tbody>
</table>
BUT…

DO WE NEED TO RECOGNIZE ALL THE DOCUMENT?
Keyword spotting

- Finding all instances of a query word that exist in a scanned document image, without fully recognizing it.
  - Document indexing and retrieval, routing, categorization...

Multi-task learning

Keyword spotting

Script identification

Existence of the keyword

Script and Writing type

CNN

Feature maps

Local features

Collapse

BLSTM

Fully Connected

global features

Collapse

Softmax

Softmax

KWS Decoder

Existence of the keyword

Script and Writing type

30/10/2020
# Results

**Script identification results (Multi-task Learning)**

<table>
<thead>
<tr>
<th>Script</th>
<th>Handwritten French</th>
<th>Handwritten Arabic</th>
<th>Printed Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handwritten French</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Handwritten Arabic</td>
<td>0%</td>
<td>99.9%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Printed Arabic</td>
<td>0%</td>
<td>0.11%</td>
<td>99.89%</td>
</tr>
</tbody>
</table>

**Keyword spotting results (100 keywords MAP)**

<table>
<thead>
<tr>
<th>System</th>
<th>Handwritten French</th>
<th>Handwritten Arabic</th>
<th>Printed Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-Task</td>
<td>89%</td>
<td>92.87%</td>
<td>98.29%</td>
</tr>
<tr>
<td>BLSTM-HMM</td>
<td>70.12%</td>
<td>84.76%</td>
<td>82.15%</td>
</tr>
<tr>
<td>HMM</td>
<td>51.9%</td>
<td>49.1%</td>
<td>61.8%</td>
</tr>
</tbody>
</table>
BUT...

IS RECOGNITION ENOUGH?
Recognition ➔ Understanding

- **READ project**
  - **Information Extraction**: Transcription + Semantic Recognition (NER)

- **Death, birth and marriage records**
Conclusion

- Whenever I say that I am a researcher in Document Analysis & Recognition, people say:

  *I thought it was a solved problem!*

Real data, realistic problem

Real data, real problem

Very difficult to reach, even for industrial solution providers
The future

- Full-page Document recognition
- There will be a new approach to OCR
  - Seq2Seq, Transformers, Graph Neural Networks...
- Labeled data is an important limitation
  - Data generation, Domain Adaptation, Few/zero-shot Learning...
- Neuro-Symbolic AI
  - Represents the causal and compositional processes behind perceptual observations
Thank you!

Yousri Kessentini
Digital research center of Sfax

yousri.kessentini@crns.rnrt.tn
http://yousri.fr.cr