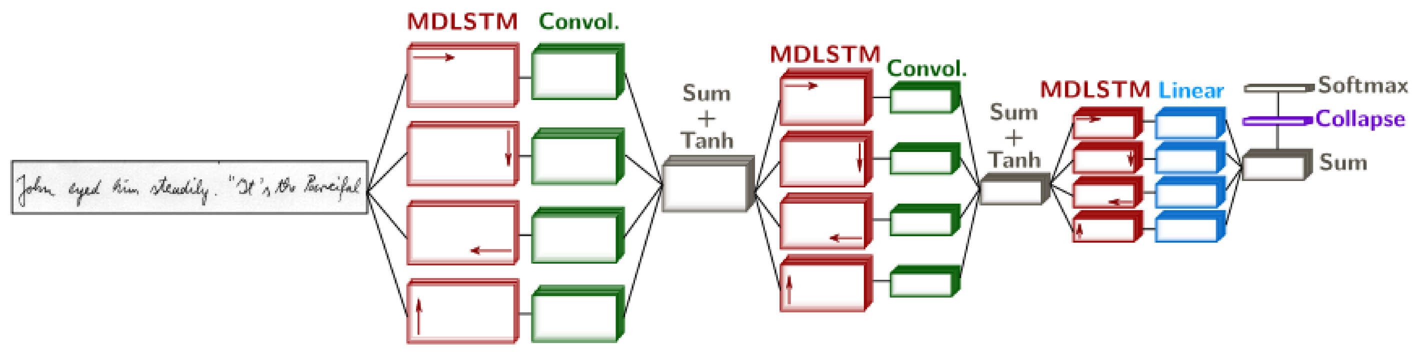
Joint Line Segmentation and Transcription for End-to-End Handwritten Paragraph Recognition

Théodore Bluche

Introduction

- Handwriting recognition evolved from isolated character recognition to word recognition with explicit segmentation to complete line recognition with implicit character segmentation.
- Nowadays handwriting recognition systems still **need cropped text lines** for both training and transcription.
- The recent works on attention models showed that it was **possible to learn to align and translate** (Bahdanau et al., 2014), or describe (Xu et al. 2015)

Handwritten Text Recognition with MDLSTM (Graves et al., 2008)



- Text line images are fed to a Multi-Dimensional LSTM layer
- Feature maps are transformed and subsampled by convolutional layers
- This is repeated several times.
- At the end, there is one feature map per character
- They are **collapsed** in the vertical direction to obtain sequences of character predictions

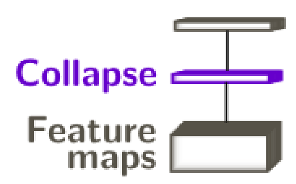
Proposed Solution

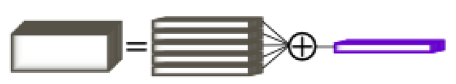
Standard Collapse

Simple vertical sum of features (all have the same importance)

Same error propagated across vertical positions

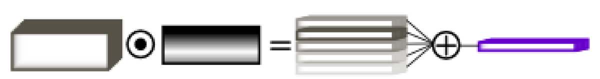
Prevents the recognition of several text lines





$$z_j = \sum_i f_{ij}$$

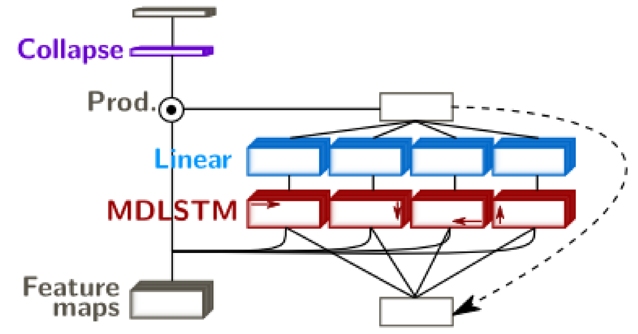
Weighted Collapse



$$z_j^{(t)} = \sum_i \omega_{ij}^{(t)} f_{ij}$$

We propose to **multiply the feature maps by a map of weights**

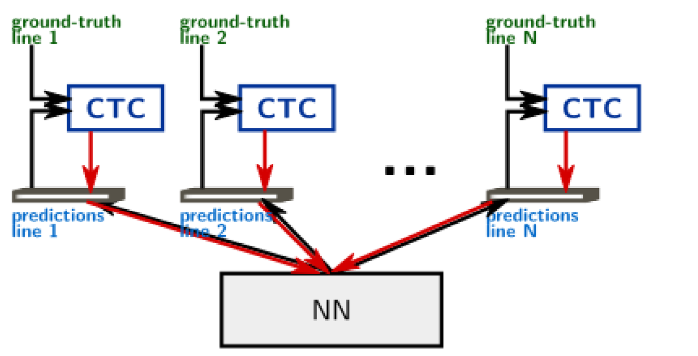
A **neural network predicts a score** for each position + softmax on columns



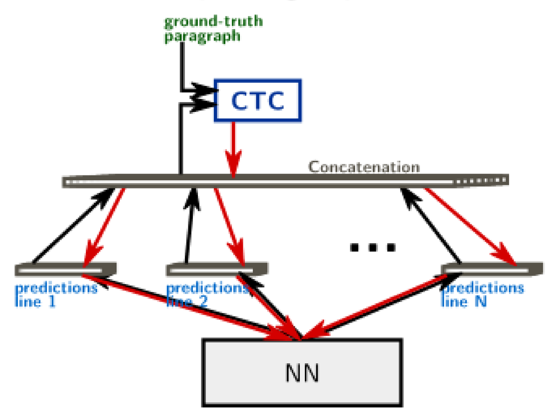
Applied several times, we can iteratively **focus on the successive text lines in a paragraph**

Model Training

Connectionist Temporal Classification (CTC) for each line if line breaks are known

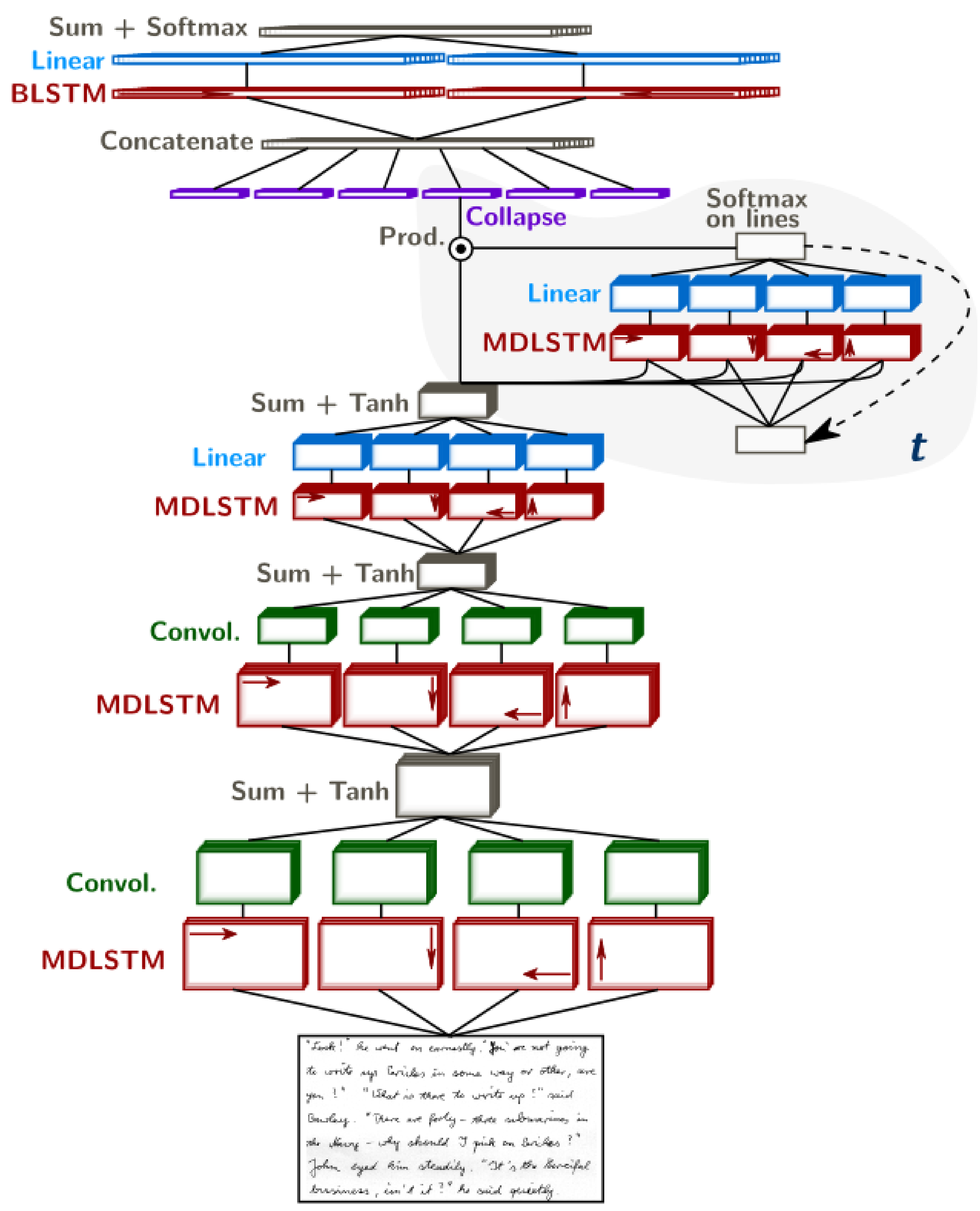


Otherwise, CTC directly on paragraphs!

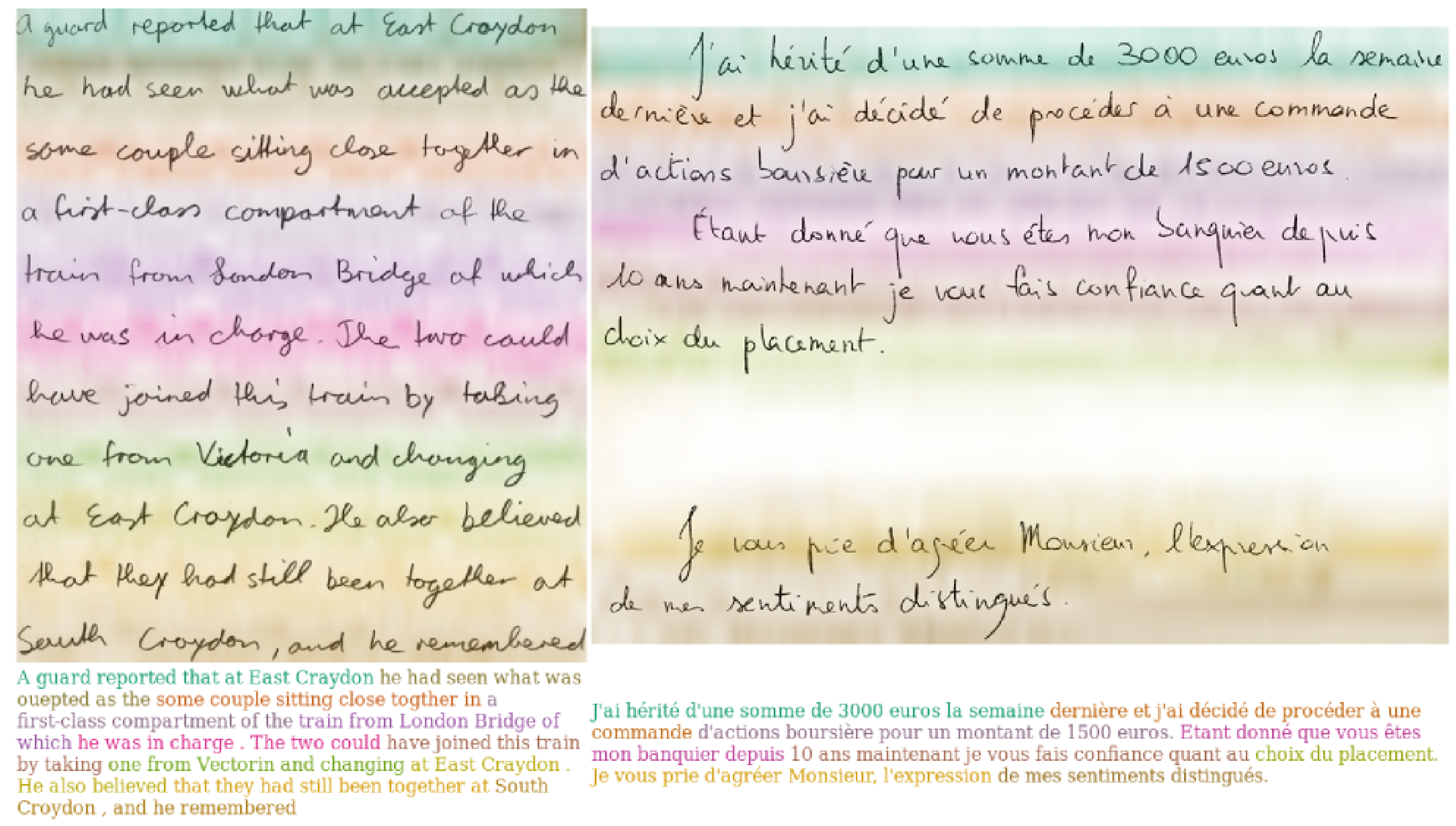


Neural Network

- We keep the **MDLSTM network** (before collapse) as an **encoder of the input paragraph image**
- The **attention network is applied iteratively** on the feature maps and performs an **implicit line segmentation**
- The obtained sequences are concatenated and fed to a **bi-directional LSTM decoder**



Qualitative Results



Quantitative Results

Outperforms line-by-line recognition with explicit line segmentation
(ground-truth positions or automatic algorithms, Char.Error.Rate%)

Database	Resolution	Line segmentation: GroundTruth	Projection	Shredding	Energy	This work
IAM	150 dpi	8.4	15.5	9.3	10.2	6.8
	300 dpi	6.6	13.8	7.5	7.9	4.9
Rimes	150 dpi	4.8	6.3	5.9	8.2	2.8
	300 dpi	3.6	5.0	4.5	6.6	2.5

Processing time comparable to segment+reco, and much faster than the "Scan, Attend and Read" model (Bluche et al., 2016)

Method		Processing time (s)
GroundTruth	(crop+reco)	0.21 ± 0.07
Shredding	(segment+crop+reco)	0.78 ± 0.26
Scan, Attend and Read	(reco)	21.2 ± 5.6
This Work	(reco)	0.62 ± 0.14

The results are **competitive with the state-of-the-art** (which uses ground-truth text-line positions)

		Rimes WER%	Rimes CER%	IAM WER%	IAM CER%
150 dpi	no language model	13.6	3.2	29.5	10.1
	with language model			16.6	6.5
300 dpi	no language model	12.6	**2.9**	24.6	7.9
	with language model			16.4	5.5
	Bluche, 2015	**11.2**	3.5	**10.9**	**4.4**
	Doetsch et al., 2014	12.9	4.3	12.2	4.7
	Kozielski et al. 2013	13.7	4.6	13.3	5.1
	Pham et al., 2014	12.3	3.3	13.6	5.1
	Messina & Kermorvant, 2014	13.3	-	19.1	-

Conclusions

We proposed a neural network for end-to-end transcription of handwritten paragraphs
An implicit line segmentation is performed by the network, which we can train directly at the paragraph level (**no need for line-level positions and annotations**)
Limitation: the attention spans the whole width of the image
=> need to refine the attention mechanism to handle arbitrary documents

tb@a2ia.com