SPAN: a Simple Predict & Align Network for Handwritten Paragraph Recognition

Denis Coquenet, Clément Chatelain, Thierry Paquet
LITIS Laboratory - EA 4108 Normandie University - University of Rouen, France

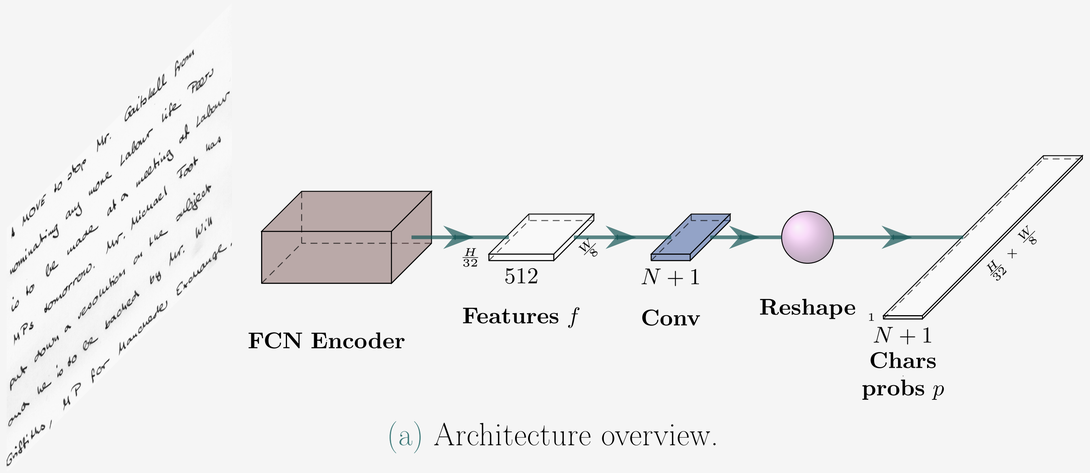
Introduction

Task: paragraph recognition

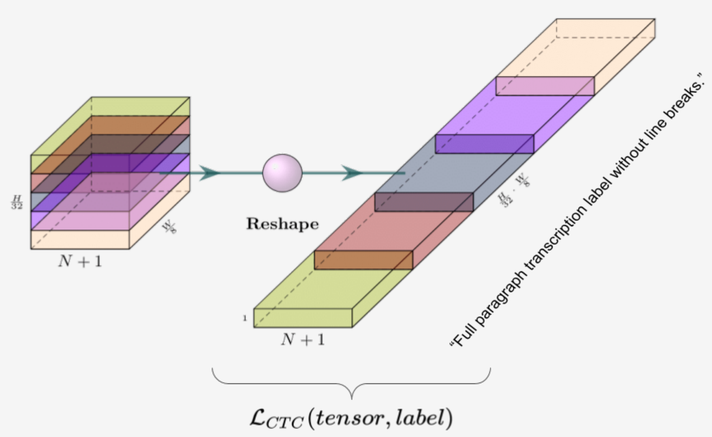
- Traditional approach in 2 steps:
  - Text line segmentation
  - Text line recognition
- Our unified end-to-end approach:
  - Whole paragraph recognition without recurrence

Our code and trained model weights are available at:
`https://github.com/FactoDeepLearning/SPAN`.

Architecture - FCN



(a) Architecture overview.



(b) Focus on the reshape operation.

- An FCN encoder extracts 2D features from the paragraph image
- Character and CTC blank label probabilities are computed for each 2D position from the features
- Rows of probabilities are concatenated to get a single 1D sequence of predictions representing the whole paragraph transcription
- The model is trained with the standard 1D CTC loss

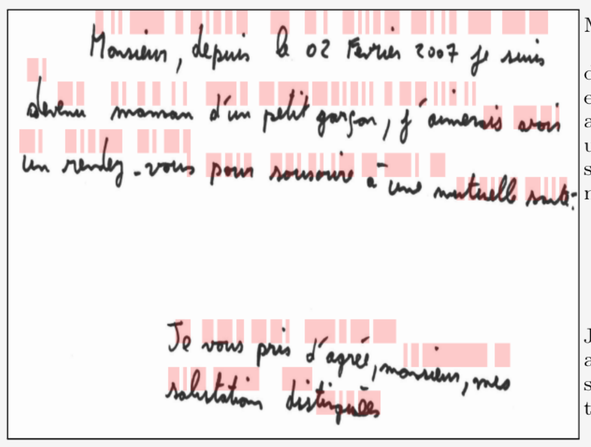
Datasets

Dataset	Level	Training	Validation	Test	Charset size
RIMES	Line	10,532	801	778	100
	Paragraph	1,400	100	100	
IAM	Line	6,482	976	2,915	79
	Paragraph	747	116	336	
READ 2016	Line	8,349	1,040	1,138	89
	Paragraph	1,584	179	197	

- Images are downscaled to 150 dpi
- The FCN Encoder is pretrained with line-level images

Prediction visualization

Network's predictions are projected on the input image, for a RIMES example.
Red: character predictions
Transparency: CTC blank label predictions

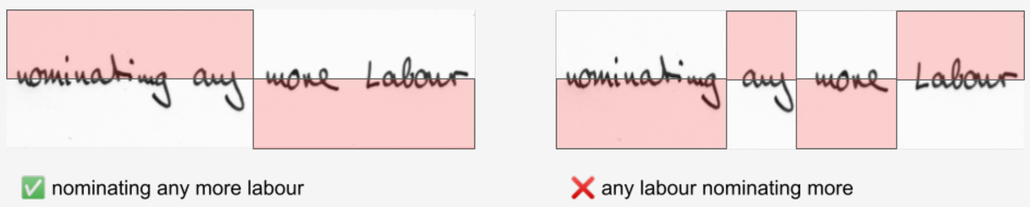


Monsieur, depuis la 02 Ferries 2007 je suis
de
enu maman d'un petit garçon, j'aimer
ais avoir
un rendez-vous
s pour sousoire à une m
ntuelle sont

Je vous pris d'agrée, m
ansieurs, mes
salutation dis
tinguées

Prediction with errors in bold:
"Monsieur, depuis l**a** 02 Fer**r**ie**s** 2007 je suis de***v***enu maman d'un petit garçon, j'aimerais avoir un rendez-vou**s**s pour sous**o**ire à une m**n**tuelle s**o**nt***é.*** Je vous pris d'agrée, m**a**nsieurs, mes salutation***s*** distinguées"

- The concatenation of the rows of probabilities implies a vertical alignment of the predictions in order to preserve the original reading order:



Results

Architecture	RIMES		IAM		READ 2016	
	CER (%)	WER (%)	CER (%)	WER (%)	CER (%)	WER (%)
[1] CNN+MDLSTM*	2.9	12.6	7.9	24.6		
[2] RPN+CNN+BLSTM	2.1	9.3	6.4	23.2		
[3] GFCN			4.7			
[4] FCN+LSTM*	**1.90**	**8.83**	**4.32**	**16.24**	**3.63**	**16.75**
This work	4.17	15.61	5.45	19.83	6.20	25.69

* with line-level attention

Our approach:

- Best path decoding
- No language model
- No lexicon constraints
- Same hyperparameters for all datasets

Conclusion

- A new approach for handwritten text recognition at paragraph level
- Deep end-to-end recurrence-free fully convolutional network
- The model is easily trained using the standard CTC loss on the whole paragraph transcription
- Competitive results on 3 datasets: IAM, RIMES and READ 2016

Github

Arxiv

References

[1] Théodore Bluche. "Joint Line Segmentation and Transcription for End-to-End Handwritten Paragraph Recognition". In: *Advances in Neural Information Processing Systems 29*. 2016, pp. 838–846.
[2] Curtis Wigington et al. "Start, Follow, Read: End-to-End Full-Page Handwriting Recognition". In: *ECCV*. Vol. 11210. Lecture Notes in Computer Science. 2018, pp. 372–388.
[3] Mohamed Yousef and Tom E. Bishop. "OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text Recognition by learning to unfold". In: *CVPR*. 2020, pp. 14698–14707.
[4] Denis Coquenet, Clément Chatelain, and Thierry Paquet. *End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network*. 2020. arXiv: 2012.03868.