Faster DAN: Multi-target Queries with Document Positional Encoding for End-to-end Handwritten Document Recognition

Denis Coquenet[1] Clément Chatelain[2,3] Thierry Paquet[2,4]
[1]Conservatoire National des Arts et Métiers, CEDRIC, Paris, France
[2]LITIS Laboratory - EA 4108 - [3]Rouen University - [4]INSA of Rouen, Rouen, France

Introduction

Handwritten Document Recognition (HDR) = recognize text & layout

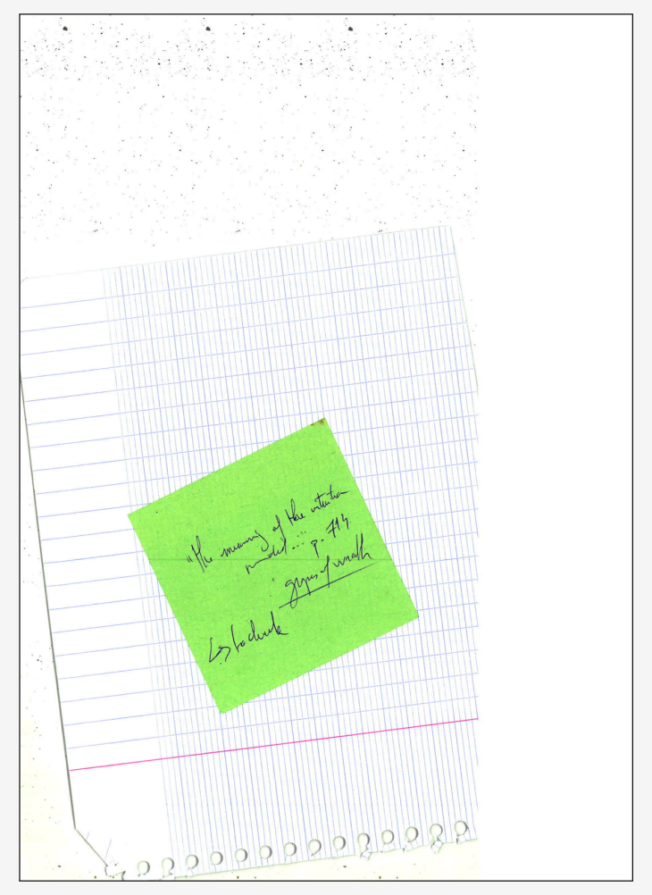
SOTA approach for end-to-end HDR: autoregressive character-level attention decoding process

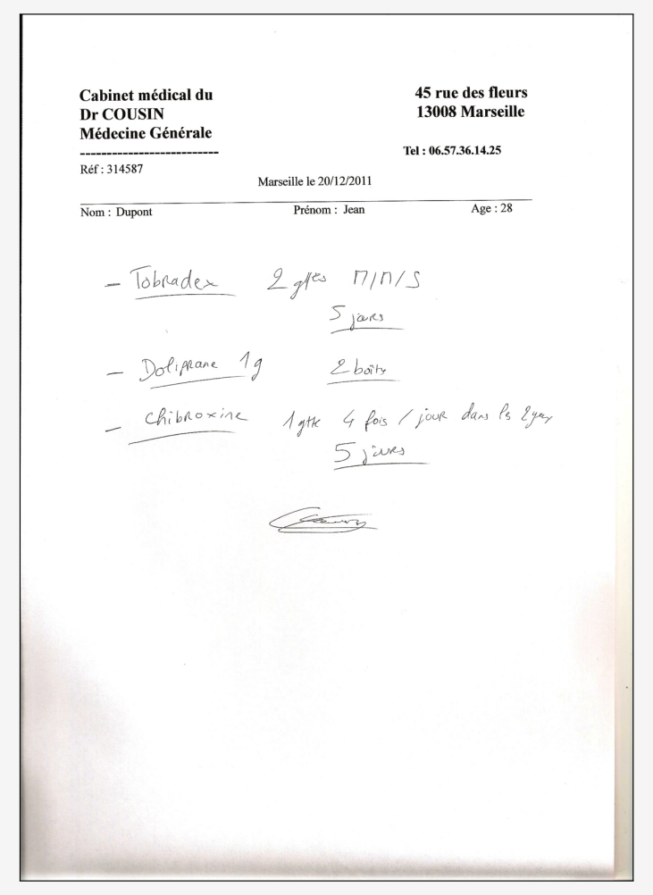
SOTA model: Document Attention Network (DAN) [1]

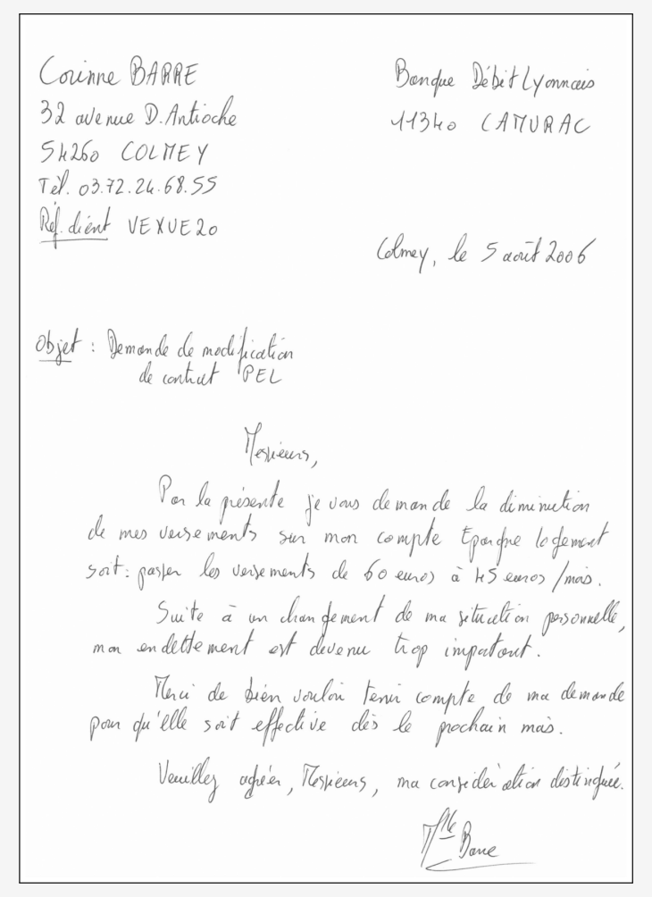
- Drawback: prediction time increases with output sequence length (∼ 1 second for 100 characters)

Goal: reducing prediction time

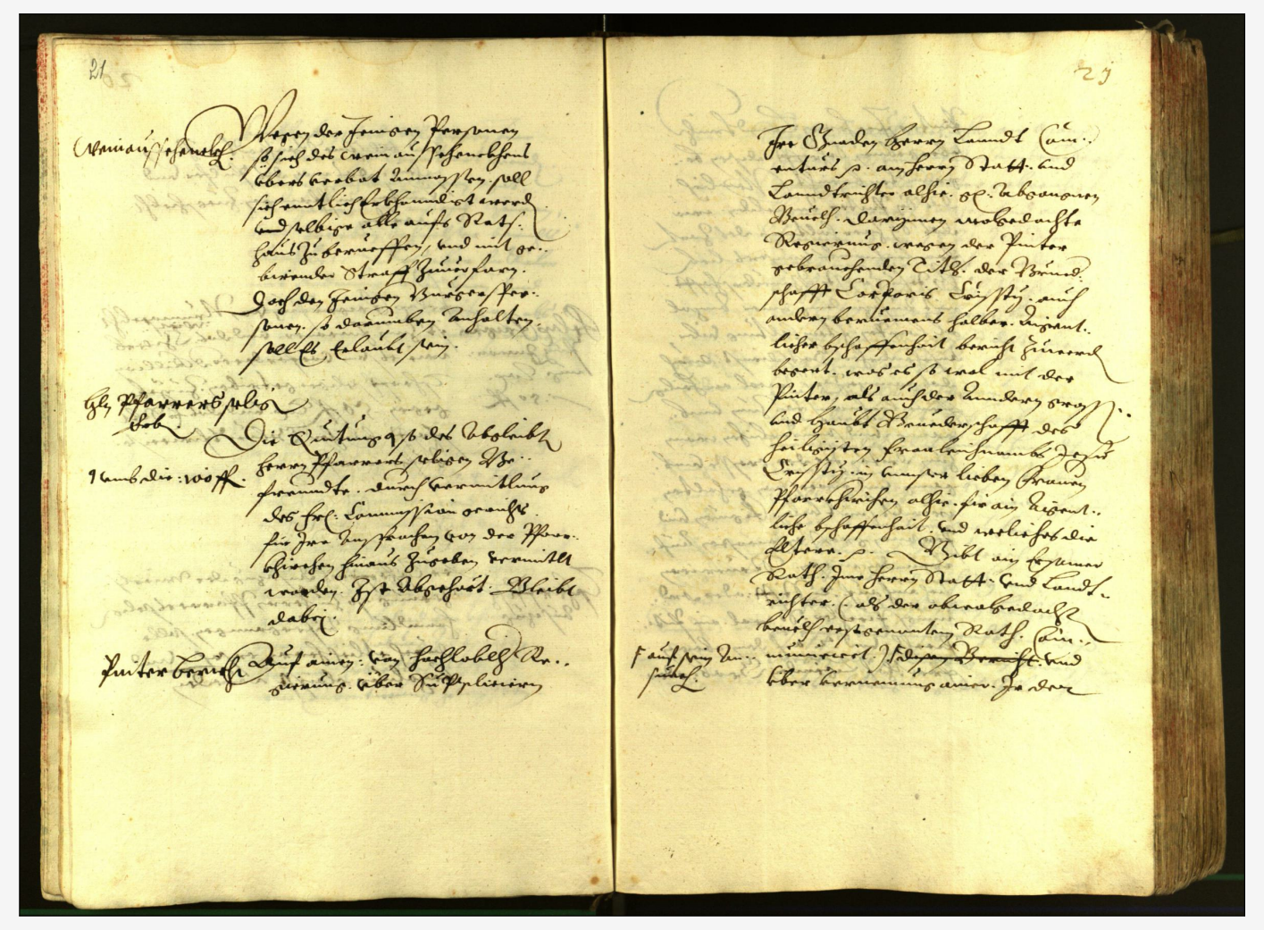
Datasets


MAURDOR C3


MAURDOR C4


RIMES 2009

READ 2016 (single-page)


READ 2016 (double-page)

Results

Architecture	READ 2016 (single-page)			READ 2016 (double-page)			RIMES 2009 (single-page)			C3	C4
	CER ↓	LOER ↓	mAP_{CER} ↑	CER ↓	LOER ↓	mAP_{CER} ↑	CER ↓	LOER ↓	mAP_{CER} ↑	CER ↓	CER ↓
DAN [1]	**3.43**	5.17	93.32	**3.70**	4.98	93.09	**4.54**	**3.82**	**93.74**	**8.62**	**8.02**
Faster DAN	3.95	**3.82**	**94.20**	3.88	**3.08**	**94.54**	6.38	4.48	91.00	8.93	9.88

CER: Character Error Rate based on string edit distance to evaluate the text recognition.
LOER: Layout Ordering Error Rate based on graph edit distance to evaluate the layout recognition.
mAP_{CER}: mean Average Precision based on a CER threshold to evaluate text & layout recognition altogether.

Prediction time (in seconds, averaged on the test set for a single document image, using a single GPU V100).

	RIMES 2009	READ 2016		MAURDOR		
	(single-page)	single-page	double-page	C3	C4	C3 & C4
DAN [1]	5.6	4.6	8.5	5.8	7.7	6.6
Faster DAN	**1.4**	**0.9**	**1.9**	**1.0**	**1.6**	**1.3**
Speed factor	x4	x5.1	x4.5	x5.8	x4.8	x5.1

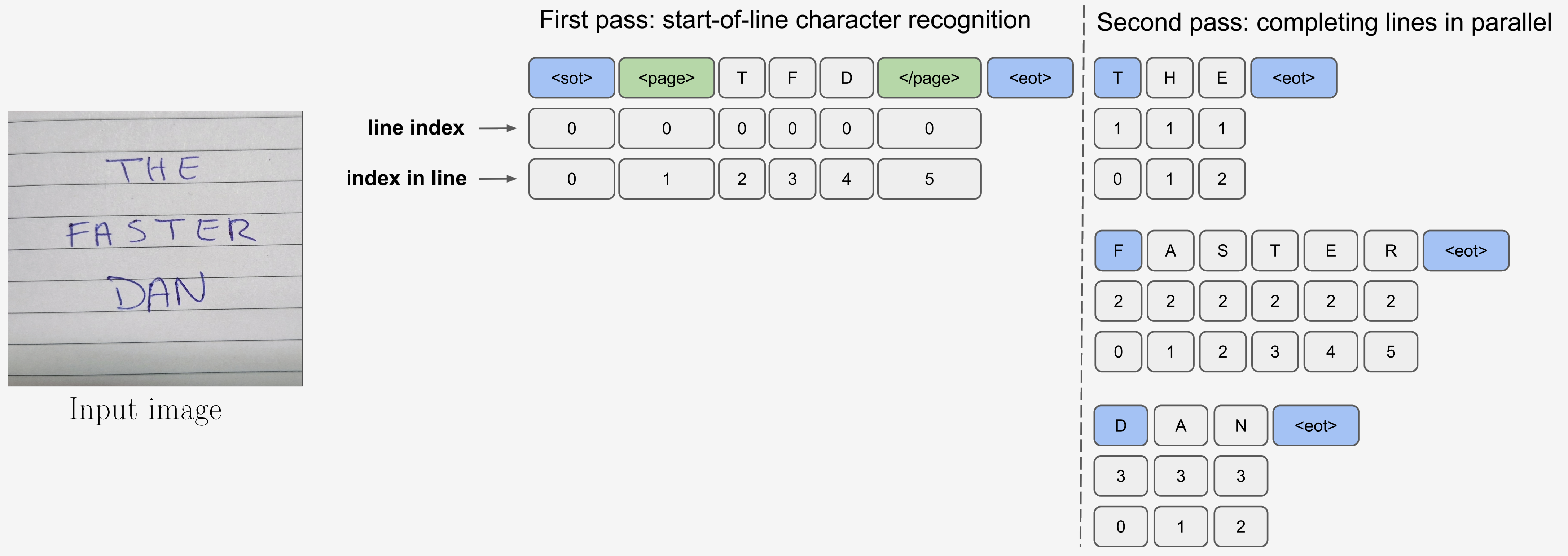
Conclusion

- Generic approach for character-level attention-based models
- Competitive results on three public datasets with the DAN architecture
- At least 4 times faster than sequential prediction process

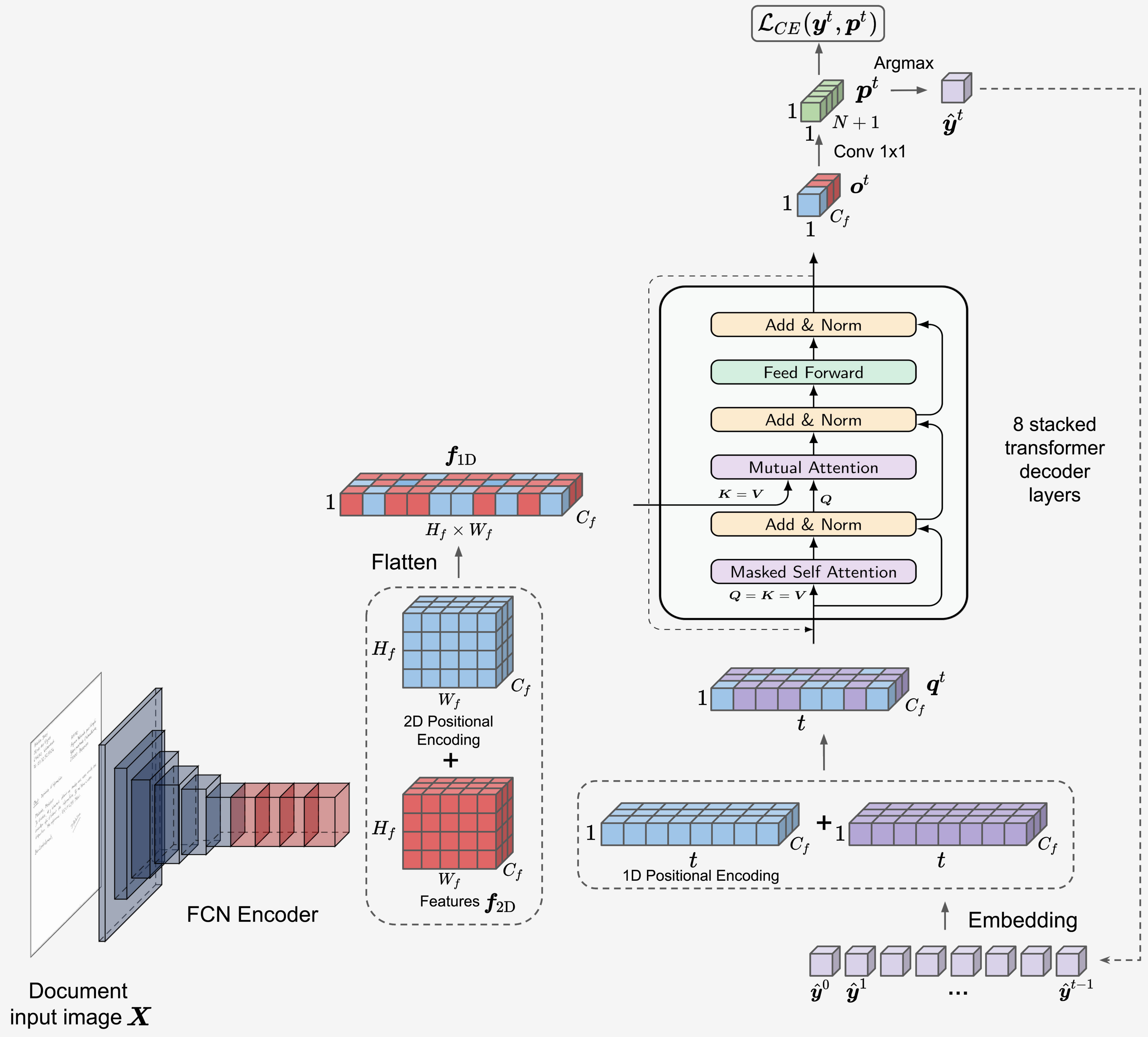
Paper, code and more !

Approach: parallelizing text line recognition

A two-step decoding process with document positional encoding:



During the first pass, the first character of each line is recognized, as well as the layout tokens (in green). The line index is set to 0.



The DAN architecture

[1] Denis Coquenet, Clément Chatelain, and Thierry Paquet. "DAN: a Segmentation-free Document Attention Network for Handwritten Document Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 45.7 (2023), pp. 8227–8243, DOI: 10.1109/TPAMI.2023.3235826.