Attention is all you need to read

Denis Coquenet

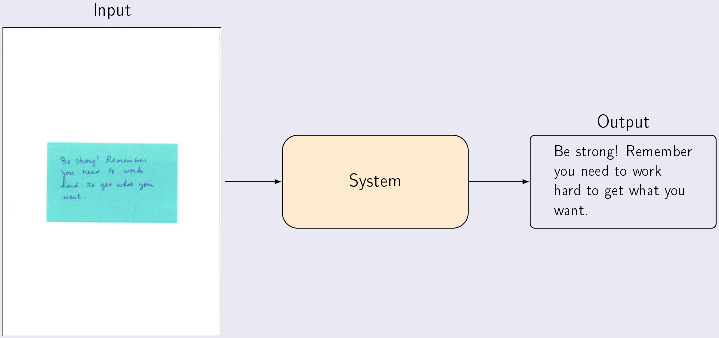
2023/09/19

Table of contents

1. Context
2. Hybrid attention for paragraph recognition
3. Autoregressive attention for document recognition
4. Speeding up with parallel attention

Handwritten Text Recognition (HTR)

An image-to-sequence problem


Input
Be strong! Remember you need to work hard to get what you want.
System
Output

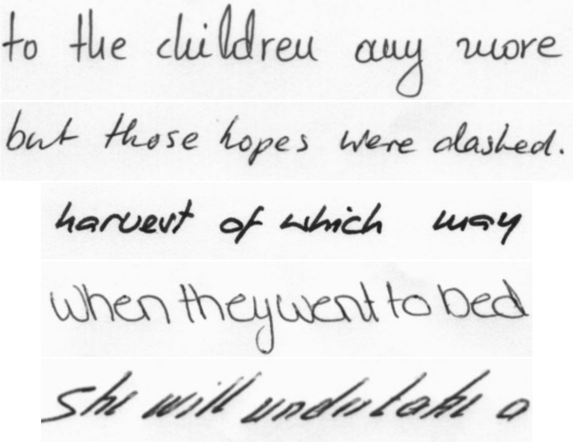
Be strong! Remember you need to work hard to get what you want.

Input: an image
Output: a sequence of characters

Challenges

A wide variety of documents

Writing styles, layout, size / resolution, background



to the children any more

but those hopes were dashed.

harvest of which way

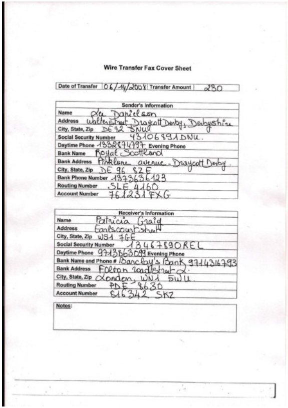
when they went to bed

She will undulate a

Challenges

A wide variety of documents

Writing styles, **layout**, size / resolution, background



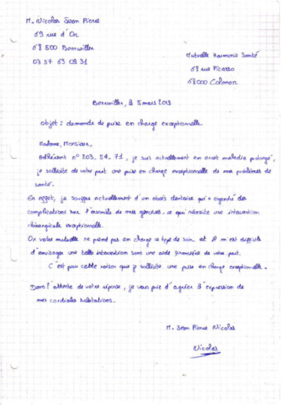
Wire Transfer Fax Cover Sheet

Date of Transfer	06/14/2008	Transfer Amount	230

Sender's Information	
Name	Lea Danielson
Address	Walterstreet, Draycott Derby, Derbyshire
City, State, Zip	DE 72 3NU.
Social Security Number	431068913DNU.
Daytime Phone	1332674797 Evening Phone
Bank Name	Royal Scottland
Bank Address	Pinklane avenue - Draycott Derby.
City, State, Zip	DE 96 8ZE
Bank Phone Number	1373636123
Routing Number	SLE 4160
Account Number	761231 FXG

Receiver's Information	
Name	Patricia Graig
Address	Earlscourtstreet
City, State, Zip	WS1 7GE
Social Security Number	134678 90REL
Daytime Phone	9713863099 Evening Phone
Bank Name and Phone #	Barclay's Bank 9714316793
Bank Address	Folton roadstreet 2.
City, State, Zip	London, WN1 5WU.
Routing Number	PDE 8630
Account Number	816342 SKZ

Notes:



M. Nicolas Jean Pierre
69 rue d'Or
68 500 Bennwiller
03 37 63 08 31

Mutuelle Harmonie Santé
68 rue Picasso
68000 Colmar

Bennwiller, le 5 mars 2013

objet : demande de prise en charge exceptionnelle

Madame, Monsieur,

Adhérent n° 203_54_71, je suis actuellement en arrêt maladie prolongé, je sollicite de votre part une prise en charge exceptionnelle de mes problèmes de santé.

En effet, je souffre actuellement d'un abcès dentaire qui a engendré des complications sur l'ensemble de mes gencives, ce qui nécessite une intervention chirurgicale exceptionnelle.

Or votre mutuelle ne prend pas en charge ce type de soin et il m'est difficile d'envisager une telle intervention sans une aide financière de votre part.

C'est pour cette raison que je sollicite une prise en charge exceptionnelle.

Dans l'attente de votre réponse, je vous prie d'agréer l'expression de mes cordiales salutations.

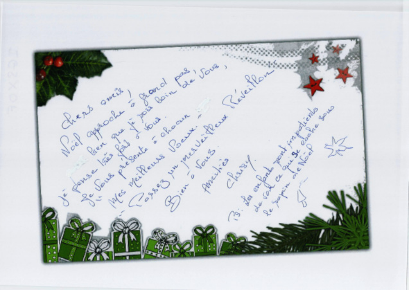
M. Jean Pierre Nicolas

Nicolas

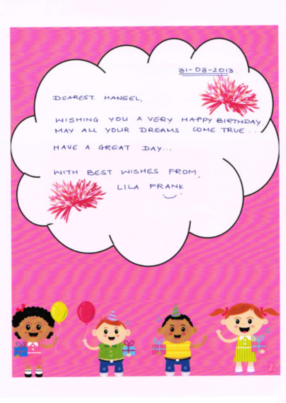
Challenges

A wide variety of documents

Writing styles, layout, size / resolution, **background**



Chers amis,
Noël approche à grand pas,
bien que je sois loin de vous,
Je pense très fort à vous.
Je vous présente à chacun
Mes meilleurs Voeux.
Passez un merveilleux Réveillon
Bien à vous
Amitiés
Chrisy.
PS: les enfants sont impatients
de voir ce qui se cache sous
le sapin de Noël



31-03-2013

DEAREST HANSEL,

WISHING YOU A VERY HAPPY BIRTHDAY.
MAY ALL YOUR DREAMS COME TRUE...

HAVE A GREAT DAY...

WITH BEST WISHES FROM,
LILA FRANK

Challenges

A wide variety of documents

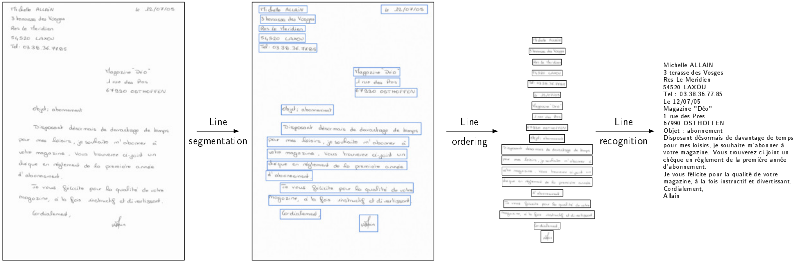
Writing styles, layout, size / resolution, background

No a priori knowledge about the document

- Number of lines
- Number of characters per line
- Reading order

The line-level sequential paradigm

- Segmentation
- Ordering
- **Recognition**

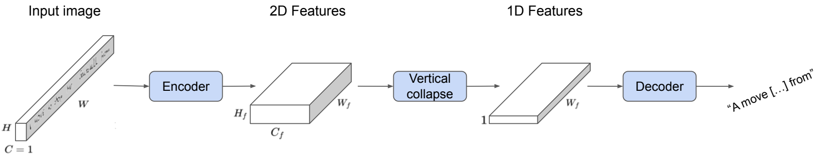


Michelle ALLAIN
3 terasse des Vosges
Res Le Meridien
54520 LAXOU
Tel : 03.38.36.77.85
Le 12/07/05
Magazine "Déo"
1 rue des Pres
67990 OSTHOFFEN
Objet : abonnement
Disposant désormais de davantage de temps pour mes loisirs, je souhaite m'abonner à votre magazine. Vous trouverez ci-joint un chèque en règlement de la première année d'abonnement.
Je vous félicite pour la qualité de votre magazine, à la fois instructif et divertissant.
Cordialement,
Allain

Related works: Recognition stage

Challenges:

- going from a 2D input image to a 1D sequence of characters
- a variable, unknown number of ordered characters to predict



Input image

Encoder

2D Features

Vertical collapse

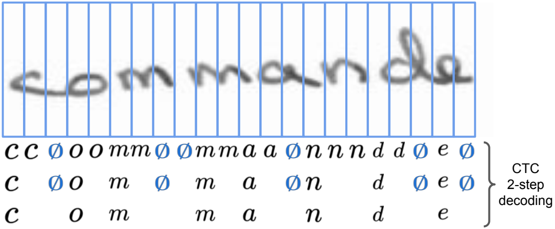
1D Features

Decoder

“A move [...] from”

Related works: Recognition stage

The Connectionist Temporal Classification (CTC) paradigm [1]

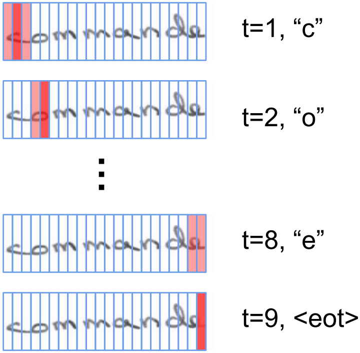

c c ∅ o o m m ∅ ∅ m m a a ∅ n n n d d ∅ e ∅
c ∅ o m ∅ m a ∅ n d ∅ e ∅
c o m m a n d e
CTC
2-step
decoding

- A frame-by-frame decision process
- Special blank token ∅
- A left-to-right constrained alignment
- ➤ CTC loss
- ➤ Limited to 1d sequences

[1] Graves *et al.*, ICML 2006

Related works: Recognition stage

The attention paradigm (at character level) [2, 3]



t=1, “c”

t=2, “o”

⋮

t=8, “e”

t=9, <eot>

- Iterative decoding process
- Implicit character segmentation
- Special end-of-transcription token <eot>
- Unconstrained attention → reading order must be learned

➤ Cross-Entropy loss

$$c^t = \sum_i \alpha_i^t f_i$$

$$\sum_i \alpha_i^t = 1$$

[2] Michael *et al.*, ICDAR 2019

Conclusion

The sequential paradigm: a mature approach... with some limitations

- Three steps treated independently
- A complex pipeline, hard to maintain
- Cumulative errors between steps
- Additional segmentation annotations
- Rule-based reading order

Goal: to overcome these limitations

Strategy: designing end-to-end HTR models step by step

➤ from line to document level

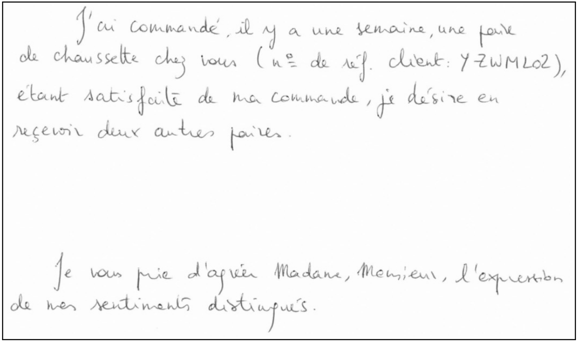
Table of contents

1. Context
2. Hybrid attention for paragraph recognition
3. Autoregressive attention for document recognition
4. Speeding up with parallel attention

Related works: Paragraph recognition

Challenges from line to paragraph recognition

- An additional vertical reading order
- Variable number of text lines
- Variable interline spacing, indent



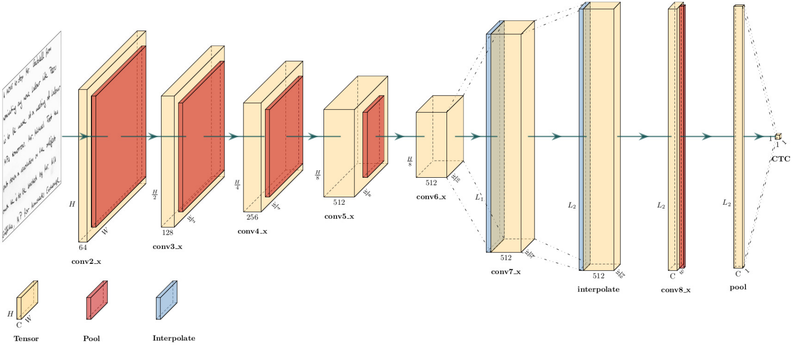
J'ai commandé, il y a une semaine, une paire
de chaussette chez vous (n° de réf. client: YZWML02),
étant satisfaite de ma commande, je désire en
recevoir deux autres paires.

Je vous prie d'agréer Madame, Monsieur, l'expression
de mes sentiments distingués.

Related works: Paragraph recognition

CTC-only approaches

- OrigamiNet [4]

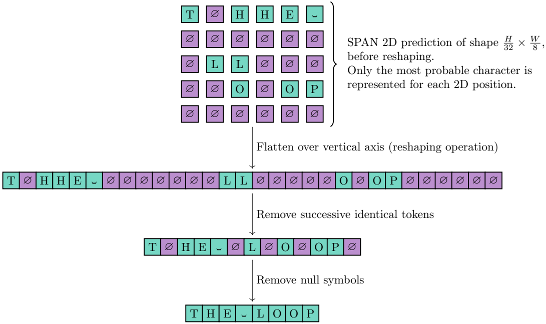


[4] Yousef *et al.*, CVPR 2020

Related works: Paragraph recognition

CTC-only approaches

- OrigamiNet [4]
- **Contribution:** Simple Predict & Align Network (SPAN) [5]



SPAN 2D prediction of shape $\frac{H}{32} \times \frac{W}{8}$, before reshaping.
Only the most probable character is represented for each 2D position.

Flatten over vertical axis (reshaping operation)

Remove successive identical tokens

Remove null symbols

[5] Coquenet *et al.*, ICDAR 2021

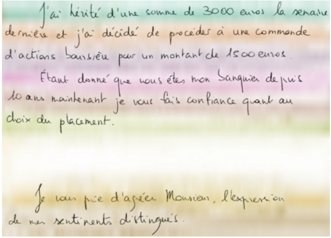
Related works: Paragraph recognition

CTC-only approaches

- OrigamiNet [4]
- **Contribution:** Simple Predict & Align Network (SPAN) [5]

Attention-based approaches

- Line-level attention [6]



J'ai hérité d'une somme de 3000 euros la semaine
dernière et j'ai décidé de procéder à une commande
d'actions boursière pour un montant de 1500 euros.
Étant donné que vous êtes mon banquier depuis
10 ans maintenant je vous fais confiance quant au
choix du placement.

Je vous prie d'agréer Monsieur, l'expression
de mes sentiments distingués.

[6] Bluche *et al.*, NIPS 2016

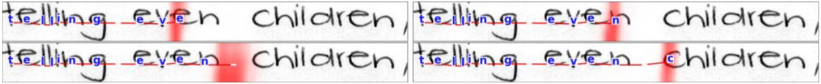
Related works: Paragraph recognition

CTC-only approaches

- OrigamiNet [4]
- **Contribution:** Simple Predict & Align Network (SPAN) [5]

Attention-based approaches

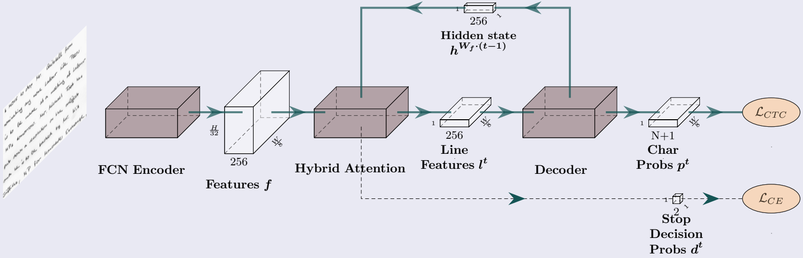
- Line-level attention [6]
- Character-level attention [7, 8]



[7] Bluche *et al.*, ICDAR 2017

Contribution: Vertical Attention Network (VAN) [9]

Overview

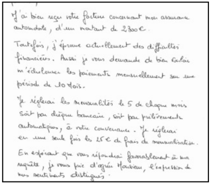


[9] Coquenet *et al.*, TPAMI 2023

Line-level vertical hybrid attention

$$\alpha_i^t = \mathrm{softmax}\left(W_\mathrm{a} \tanh(W_\mathrm{f} f'_i + W_\mathrm{j} j_{t,i} + W_h h_{W_f(t-1)})\right)$$

Datasets

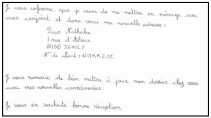


J'ai bien reçu votre facture concernant mon assurance
automobile, d'un montant de 2300€.

Toutefois, j'éprouve actuellement des difficultés
financières. Aussi je vous demande de bien vouloir
m'échelonner les paiements mensuellement sur une
période de 10 mois.

Je règlerai les mensualités le 5 de chaque mois
soit par chèque bancaire, soit par prélèvements
automatiques, à votre convenance. Je règlerai
en une seule fois les 25€ de frais de mensualisation.

En espérant que vous répondrai favorablement à ma
requête, je vous prie d'agréer Monsieur, l'expression de
mes sentiments distingués.

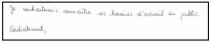


Je vous informe que je viens de me mettre en ménage avec
mon conjoint et donc voici ma nouvelle adresse :

Paco Nathalie
1 rue d'Alsace
88150 IGNEY
N° de client : NMRRZ02

Je vous remercie de bien mettre à jour mon dossier chez vous
avec mes nouvelles coordonnées.

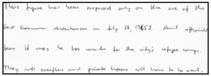
Je vous en souhaite bonne réception.



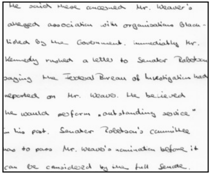
Je souhaiterais connaître vos horaires d'accueil au public.

Cordialement,

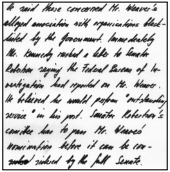
RIMES 2011



This figure has been surpassed only on the eve of the
East German disturbances on July 17, 1953. And officials
fear it may be too much for the city's refugee camps.
They will overflow and private homes will have to be used.

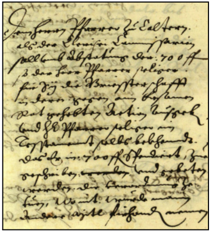


He said these concerned Mr. Weaver's
alleged association with organisations Black-
listed by the Government. Immediatly Mr.
Kennedy rushed a letter to Senator Robertson
saying the Federal Bureau of Investigation had
reported on Mr. Weaver. He believed
he would perform „outstanding service"
in his post. Senator Robertson's committee
has to pass Mr. Weaver's nomination before it
can be considered by the full Senate.

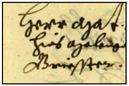


He said these concerned Mr. Weaver's
alleged association with organisations black-
listed by the Government. Immedeately
Mr. Kennedy rushed a letter to Senator
Robertson saying the Federal Bureau of In-
vestigation had reported on Mr. Weaver.
He believed he would perform "outstanding
service" in his post. Senator Robertson's
comittee has to pass Mr. Weaveó
nomination before it can be con-
~~[illegible]~~ sidered by the full Senate.

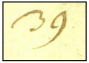
IAM



[illegible] Pfarrer zu [illegible]
[illegible]
[illegible] 700 fl
[illegible] Pfarrer [illegible]
[illegible]
[illegible]
[illegible]
[illegible] Pfarrer [illegible]
[illegible]
[illegible] 700 fl [illegible]
[illegible]
[illegible]
[illegible]
[illegible]



[illegible]
[illegible]
[illegible]



39

READ 2016

Paragraph-level recognition results

Paragraph-level state-of-the-art approaches, without language model, external data, nor lexicon constraints.

Architecture	IAM		RIMES 2011		READ 2016		# Param.
	CER (%) test	WER (%) test	CER (%) test	WER (%) test	CER (%) test	WER (%) test	
Best line-level approach	4.87[1]		2.3[2]	9.6[2]	4.66[1]		
[7] CNN+MDLSTM[b]	16.2						
[6] CNN+MDLSTM[a]	7.9	24.6	2.9	12.6			
[8] CNN+Transformer[b]	6.7						27.8 M
[5] SPAN (FCN)	5.45	19.83	4.17	15.61	6.20	25.69	19.2 M
[4] OrigamiNet (GFCN)	4.7						16.4 M
[9] VAN (FCN+LSTM)[a]	**4.45**	**14.55**	**1.91**	**6.72**	**3.59**	**13.94**	2.7 M

[1] Results from [2] CNN+BLSTM[b].

[2] Results from [10] CNN+BLSTM.

[a] With line-level attention.

[b] With character-level attention.

VAN demonstration

`https://youtu.be/OXi1birmbuw`

Conclusion

Attention is powerful but:

- Attention mechanisms → slower convergence
  ➤ vertical attention (1D) + pre-training
- Hybrid attention
  ➤ recurrent training (OK for lines, KO for chars)

Bridging the gap between line-level and paragraph-level approaches...

- State-of-the-art results on RIMES 2011, IAM and READ 2016
- Able to deal with slightly inclined lines

... but still the same limitations, inherent to the sequential paradigm

➤ Rethinking the paradigm

Table of contents

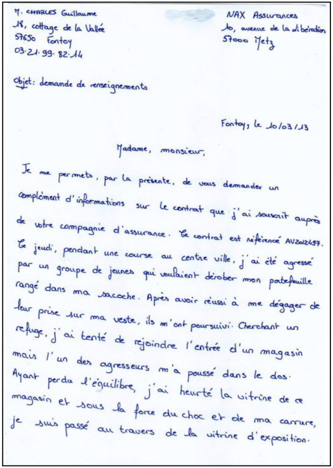
1 Context

2 Hybrid attention for paragraph recognition

3 Autoregressive attention for document recognition

4 Speeding up with parallel attention

HTR at document level



M. CHARLES Guillaume
18, cottage de la Vallée
57650 Fontoy
03.21.99.82.14

NAX Assurances
10, avenue de la libération
57000 Metz

Objet: demande de renseignements

Fontoy, le 10/03/13

Madame, monsieur,

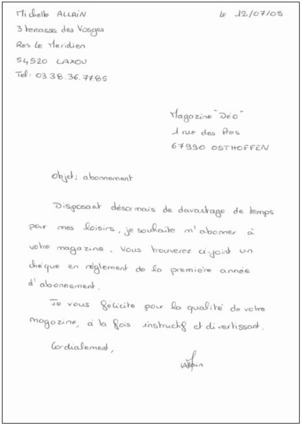
Je me permets, par la présente, de vous demander un complément d'informations sur le contrat que j'ai souscrit auprès de votre compagnie d'assurance. Ce contrat est référencé AV2012497. Ce jeudi, pendant une course au centre ville, j'ai été agressé par un groupe de jeunes qui voulaient dérober mon portefeuille rangé dans ma sacoche. Après avoir réussi à me dégager de leur prise sur ma veste, ils m'ont poursuivi. Cherchant un refuge, j'ai tenté de rejoindre l'entrée d'un magasin mais l'un des agresseurs m'a poussé dans le dos. Ayant perdu l'équilibre, j'ai heurté la vitrine de ce magasin et sous la force du choc et de ma carrure, je suis passé au travers de la vitrine d'exposition.

Challenges from paragraph to document

- Layout-dependent reading order
- Larger input images and output sequences
  - ➤ GPU constraints
  - ➤ More complex attention

Handwritten Document Recognition (HDR)

Goal: joint recognition of both text and layout from whole documents



Michelle ALLAIN le 12/07/05
3 terasse des Vosges
Res Le Meridien
54520 LAXOU
Tel: 03.38.36.77.85

Magazine "Déo"
1 rue des Pres
67990 OSTHOFFEN

objet: abonnement

Disposant désormais de davantage de temps pour mes loisirs, je souhaite m'abonner à votre magazine. Vous trouverez ci-joint un chèque en réglement de la première année d'abonnement.

Je vous félicite pour la qualité de votre magazine, à la fois instructif et divertissant.

Cordialement,

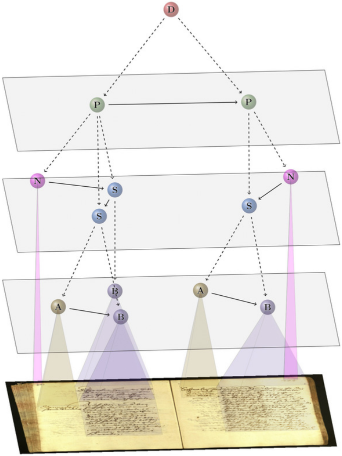
Allain

Handwritten Document → Recognition

Michelle ALLAIN
3 terasse des Vosges
Res Le Meridien
54520 LAXOU
Tel : 03.38.36.77.85
Le 12/07/05
Magazine "Déo"
1 rue des Pres
67990 OSTHOFFEN
Objet : abonnement
Disposant désormais de davantage de temps
pour mes loisirs, je souhaite m'abonner à
votre magazine. Vous trouverez ci-joint un
chèque en règlement de la première année
d'abonnement.
Je vous félicite pour la qualité de votre
magazine, à la fois instructif et divertissant.
Cordialement,
Allain

🟫	Sender Coordinates
🟦	Recipient Coordinates
🟩	Place & Date
⬜	Object
🟪	Body
🟧	Signature

How to encode both text and layout ?



```
<document>
  <page>
    <page_number>
      204
    </page_number>
    <section>
      <body>
        Schgrafer, [...] gehalt.
      </body>
    </section>
    <section>
      <annotation>
        Genneral [...] Raitūng
      </annotation>
      <body>
        Aūf den: [...] werden,
      </body>
    </section>
  </page>
  <page>
    <page_number>
      204
    </page_number>
    <section>
      <annotation>
        Schmalz. [...] bet:
      </annotation>
      <body>
        Verer [...] dar¬
      </body>
    </section>
  </page>
</document>
```

- XML paradigm

How to evaluate the performance ?

Evaluate the text recognition

- CER / WER

➤ Normalized edit distance between sequences of characters / words

Prediction: "<A><B>HTR</B>2<B>HDR</B></A>"
Metric computed on: "HTR2HDR"

How to evaluate the performance ?

Evaluate the text recognition

- CER / WER

Evaluate the layout recognition

- LOER (Layout Ordering Error Rate)

➤ Normalized edit distance between graphs

Prediction: "<A><B>HTR</B>2<B>HDR</B></A>"
Metric computed on: "<A><B></B><B></B></A>"

How to evaluate the performance ?

Evaluate the text recognition

- CER / WER

Evaluate the layout recognition

- LOER (Layout Ordering Error Rate)

⚠ **Not sufficient:**

Ground truth: "<A><B>HTR</B>2<B>HDR</B></A>"
Prediction: "<A><B></B><B></B></A>HTR2HDR"

LOER = 0%   CER = 0%

How to evaluate the performance ?

Evaluate the text recognition

- CER / WER

Evaluate the layout recognition

- LOER (Layout Ordering Error Rate)

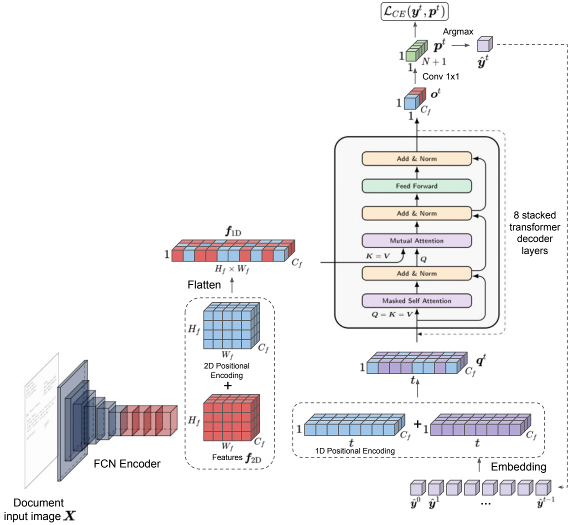
Evaluate text and layout recognition altogether

- mAP_{CER}

➤ Area under the precision / recall curve

Prediction: "<A><B>HTR</B>2<B>HDR</B></A>"
Metric computed on: "HTR2HDR", "HTR", "HDR"

Document Attention Network (DAN) [11]



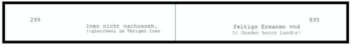
$$\boldsymbol{c}^t = \underbrace{\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)}_{\alpha^t}\boldsymbol{V}$$

→ Teacher forcing

[11] Coquenet *et al.*, TPAMI 2023

DAN - Training strategy

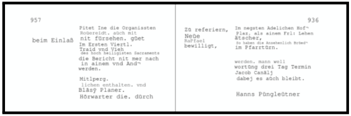
- Pre-training on synthetic text line images.
- Curriculum learning with synthetic documents:



299
Inen nicht nachzeseh.
(:gleichwol im Vbrigen Iren

895
feltigs Ermanen vnd
Ir Gnaden herrn Landts¬

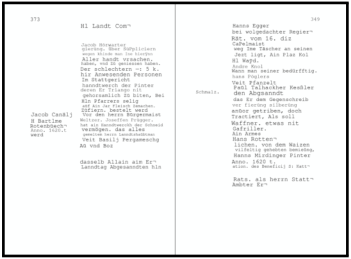
(a) $l = 3$.



957
beim Einlaß
Pitet Ine die Organissten
Roůereidt. aůch mit
nit fůrsehen. gůet
Im Ersten Viertl.
Traid vnd Vieh
des hoch heiligisten Sacraments
die Bericht nit mer nach
in ainem vnd And¬
werden.
Mitlperg.
lichen enthalten. vnd
Bläsÿ Planer.
Hörwarter die. důrch

936
Zů referiern,
Neůe
RaPfael
bewilligt,
Im negsten Adelichen Hof¬
Plaz. als ainem Frl: Lehen
ätscher,
So haben die Ansehenlich Brůed¬
im Pfarrtůrn.
werden. mann well
wortůng drei Tag Termin
Jacob Canälj
dabej es aůch bleibt.
Hanns Půngleůtner

(b) $l = 15$.

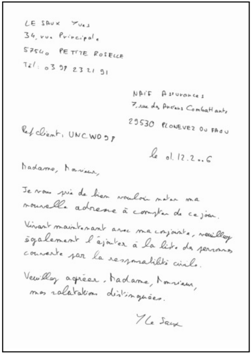


373
Hl Landt Com¬
Jacob Hörwarter
gierůng. Über SůPpliciern
wegen khinde man Ine hierÿnn
Aller handt vrsachen.
haben, vnd Zů geniessen haben.
Der schlechtern –: 5 k.
hir Anwesenden Personen
Im Stattgericht
hanndtwerch der Pinter
deren Er Triango nit
gehorsamlich Zů biten, Bei
Hln Pfarrers selig
aůf Ain Jar Fleisch Zemachen.
Zůfiern. bestelt werd
Jacob Canälj
H Bartlme
Rotenbůech¬
Anno. 1620.t
werd
Vor den herrn Bůrgermaist
Weltzer. Joseffen Průgger.
hat ain Hanndtwerckh der Schneid
vermögen. das alles
gemeltem herrn Lanndtshaůbtman
Veit Basilj Pergameschg
Aů vnd Boz
dasselb Allain aim Er¬
Lanndtag Abgesanndten hln

349
Hanns Egger
bei wolgedachter Regier¬
Rät. vom 16. diz
CaPelmaist
weg Ime Täscher an seinen
Jezt ligt, Ain Plaz Kol
Hl Waÿd.
Andre Knol
Wann man seiner bedůrfftig.
hans Pöglers
Veit Pfanzelt
Schmalz.
Paůl Talhackher Keßler
den Abgsanndt
das Er dem Gegenschreib
ver fierůng silberůng
anůor getriben, doch
Tractiert, Als soll
Waffner. etwas nit
Gafriller.
Ain Armes
Hans Rotten¬
lichen. von dem Waizen
vilfeltig gehebten bemieůng,
Hanns Mirdinger Pinter
Anno. 1620 t.
ation. des Beneficij S: Katt¬
Rats. als herrn Statt¬
Ambter Er¬

(c) $l = l_{\max} = 30$ (end of curriculum stage, no crop).

Datasets



LE SAUX Yves
34, rue Principale
57540 PETITE ROSELLE
Tél : 03 97 23 21 91

NAIF Assurances
7, rue des Anciens Combattants
29530 PLONEVEZ DU FAOU

Réf client : UNCW097

le 01.12.2006

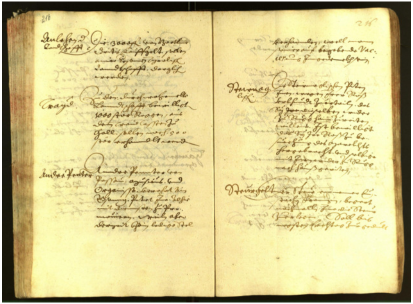
Madame, Monsieur,

Je vous prie de bien vouloir noter ma nouvelle adresse à compter de ce jour.

Vivant maintenant avec ma conjointe, veuillez également l'ajouter à la liste des personnes couverte par la responsabilité civile.

Veuillez agréer, Madame, Monsieur, mes salutations distinguées.

Y Le Saux



Dataset	Level	Training	Validation	Test	# char tokens	# layout tokens
RIMES 2009 [12]	Page	1,050	100	100	108	14
READ 2016 [13]	Page	350	50	50	89	10
	Double page	169	24	24		

DAN results on the RIMES dataset

⚠ Metrics do not take into account the segmentation step

Dataset	Approach	CER (%) ↓	WER (%) ↓	LOER (%) ↓	mAP_{CER} (%) ↑
RIMES 2011	**Line level**				
	[9] FCN	3.04	8.32	✗	✗
	[10] CNN+BLSTM[a]	**2.3**	9.6	✗	✗
	[11] DAN (FCN+transformer)[c]	2.63	**6.78**	✗	✗
	Paragraph level				
	[5] SPAN (FCN)	4.17	15.61	✗	✗
	[6] CNN+MDLSTM[b]	2.9	12.6	✗	✗
	[9] VAN (FCN+LSTM)[b]	1.91	6.72	✗	✗
	[11] DAN (FCN+transformer)[c]	**1.82**	**5.03**	✗	✗
RIMES 2009	**Paragraph level**				
	[11] DAN (FCN+transformer)[c]	5.46	13.04	✗	✗
	Page level				
	[11] DAN (FCN+transformer)[c]	4.54	11.85	3.82	93.74

[a] This work uses a slightly different split (10,203 for training, 1,130 for validation and 778 for test).

[b] with line-level attention.

[c] with character-level attention.

DAN results on the READ 2016 dataset

⚠ Metrics do not take into account the segmentation step

Approach	CER (%) ↓	WER (%) ↓	LOER (%) ↓	mAP_{CER} (%) ↑
Line level				
[2] CNN+BLSTM[a]	4.66	✗	✗	✗
[13] CNN+RNN	5.1	21.1	✗	✗
[9] VAN (FCN+LSTM)[b]	**4.10**	**16.29**	✗	✗
[11] DAN (FCN+transformer)[a]	**4.10**	17.64	✗	✗
Paragraph level				
[5] SPAN (FCN)	6.20	25.69	✗	✗
[9] VAN (FCN+LSTM)[b]	3.59	13.94	✗	✗
[11] DAN (FCN+transformer)[a]	**3.22**	**13.63**	✗	✗
Single-page level				
[11] DAN (FCN+transformer)[a]	3.53	13.33	5.94	92.57
Double-page level				
[11] DAN (FCN+transformer)[a]	3.69	14.20	4.60	93.92

[a] with character-level attention.

[b] with line-level attention.

DAN demonstration

`https://youtu.be/HrrUsQfW66E`

Conclusion

DAN: the first end-to-end model for HDR

- Structured output sequence
- No need for any physical segmentation annotation
- Can follow the slant of the lines (character-level attention)

Line-level / paragraph-level limitations

- ~~Three steps treated independently~~
- ~~A complex pipeline, hard to maintain~~
- ~~Cumulative errors between steps~~
- ~~Additional segmentation annotations~~
- ~~Rule-based reading order~~

Drawback: prediction times grow with the character sequence

Table of contents

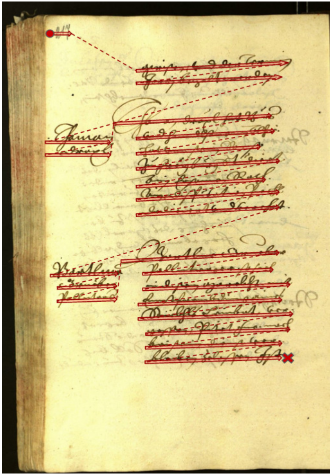
1 Context

2 Hybrid attention for paragraph recognition

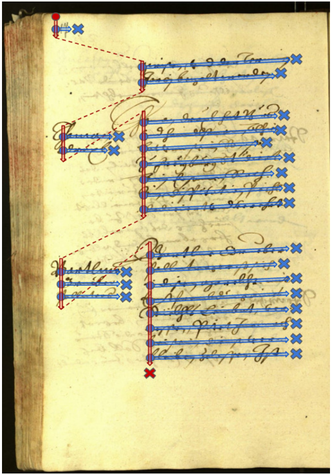
3 Autoregressive attention for document recognition

4 Speeding up with parallel attention

Faster DAN: parallelizing text line recognition [14]



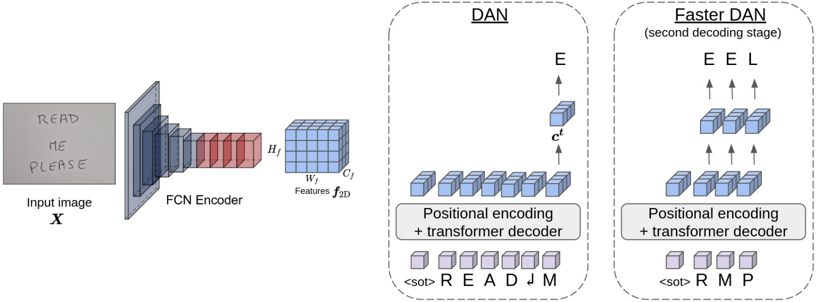
(a) DAN



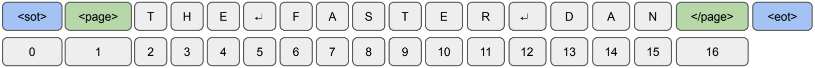
(b) Faster DAN

[14] Coquenet *et al.*, ICDAR 2023

Faster DAN - Multi-target queries

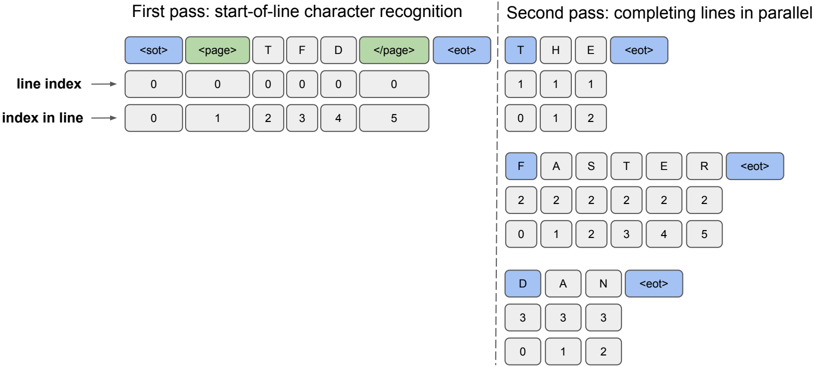


Faster DAN - Positional encoding



<sot>	<page>	T	H	E	↵	F	A	S	T	E	R	↵	D	A	N	</page>	<eot>
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	

(a) DAN single-pass prediction process



First pass: start-of-line character recognition

	<sot>	<page>	T	F	D	</page>	<eot>
line index →	0	0	0	0	0	0	
index in line →	0	1	2	3	4	5	

Second pass: completing lines in parallel

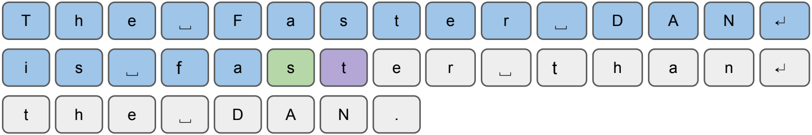
T	H	E	<eot>
1	1	1	
0	1	2	

F	A	S	T	E	R	<eot>
2	2	2	2	2	2	
0	1	2	3	4	5	

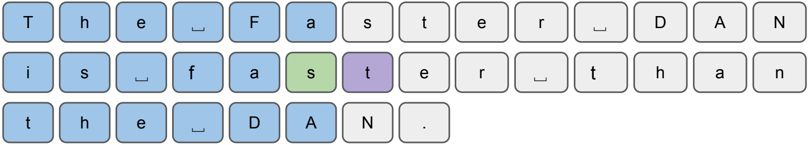
D	A	N	<eot>
3	3	3	
0	1	2	

(b) Faster DAN two-pass prediction process

Faster DAN - Context



(a) Context used by the DAN



(b) Context used by the Faster DAN

Results

Architecture	READ 2016 (single-page)				READ 2016 (double-page)			
	CER ↓	WER ↓	LOER ↓	mAP_{CER} ↑	CER ↓	WER ↓	LOER ↓	mAP_{CER} ↑
DAN [11]	**3.43**	**13.05**	5.17	93.32	**3.70**	**14.15**	4.98	93.09
Faster DAN [14]	3.95	14.06	**3.82**	**94.20**	3.88	14.97	**3.08**	**94.54**

Architecture	RIMES 2009			
	CER ↓	WER ↓	LOER ↓	mAP_{CER} ↑
DAN [11]	**4.54**	**11.85**	**3.82**	**93.74**
Faster DAN [14]	6.38	13.69	4.48	91.00

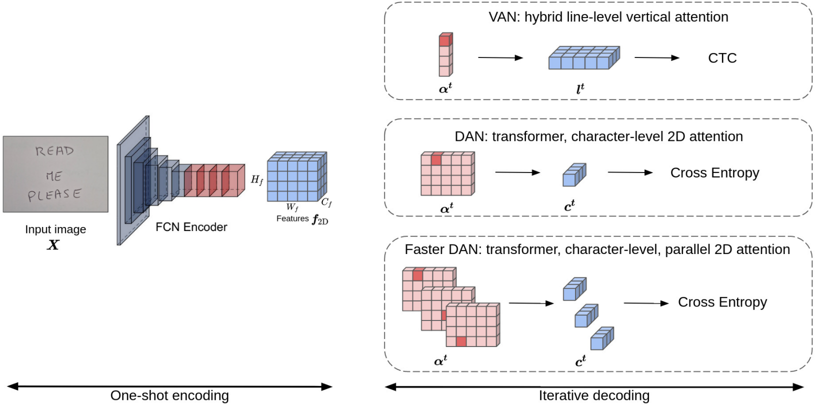
Prediction times

	RIMES 2009	READ 2016		MAURDOR		
		single-page	double-page	C3	C4	C3 & C4
Dataset details (averaged for a document on the test set)						
width (px)	1,235	1,190	2,380	1,336	1,240	1,297
height (px)	1,751	1,755	1,755	1,658	1,754	1,697
# chars	578	528	1,062	481	706	575
# lines	18	23	47	16	22	18
# chars / line	31	22	22	30	31	30
# layout tokens	11	15	30	0	0	0
Prediction times (in seconds)						
DAN [11]	5.6	4.6	8.5	5.8	7.7	6.6
Faster DAN [14]	**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

Faster DAN demonstration

https://youtu.be/_pBsO2W8XRE

VAN vs DAN vs Faster DAN



Input image $\boldsymbol{X}$

FCN Encoder

H_f W_f C_f

Features $\boldsymbol{f}_{2D}$

VAN: hybrid line-level vertical attention

$\boldsymbol{\alpha}^t$ → $\boldsymbol{l}^t$ → CTC

DAN: transformer, character-level 2D attention

$\boldsymbol{\alpha}^t$ → $\boldsymbol{c}^t$ → Cross Entropy

Faster DAN: transformer, character-level, parallel 2D attention

$\boldsymbol{\alpha}^t$ → $\boldsymbol{c}^t$ → Cross Entropy

One-shot encoding

Iterative decoding

General conclusion

Attention for reading systems

Line → Paragraph → Document

➤ From text recognition to reading

Perspectives

Recognizing more:

- Heterogeneous documents (layout)
- Multilingual documents
- Combining HDR with other tasks: Named Entity Recognition, Mathematical Expression Recognition, Table Recognition

Thank you for your attention

References I

[1] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: *International Conference on Machine Learning (ICML)*. Vol. 148. 2006, pp. 369–376.

[2] Johannes Michael, Roger Labahn, Tobias Grüning, and Jochen Zöllner. "Evaluating Sequence-to-Sequence Models for Handwritten Text Recognition". In: *International Conference on Document Analysis and Recognition (ICDAR)*. 2019, pp. 1286–1293.

[3] Christoph Wick, Jochen Zöllner, and Tobias Grüning. "Transformer for Handwritten Text Recognition Using Bidirectional Post-decoding". In: *16th International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 12823. 2021, pp. 112–126.

[4] Mohamed Yousef and Tom E. Bishop. "OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text Recognition by learning to unfold". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 14698–14707.

References II

[5] Denis Coquenet, Clément Chatelain, and Thierry Paquet. "SPAN: a Simple Predict & Align Network for Handwritten Paragraph Recognition". In: *International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 12823. 2021, pp. 70–84.

[6] Théodore Bluche. "Joint Line Segmentation and Transcription for End-to-End Handwritten Paragraph Recognition". In: *Advances in Neural Information Processing Systems 29 (NIPS)*. 2016, pp. 838–846.

[7] Théodore Bluche, Jérôme Louradour, and Ronaldo O. Messina. "Scan, Attend and Read: End-to-End Handwritten Paragraph Recognition with MDLSTM Attention". In: *International Conference on Document Analysis and Recognition (ICDAR)*. 2017, pp. 1050–1055.

[8] Sumeet S. Singh and Sergey Karayev. "Full Page Handwriting Recognition via Image to Sequence Extraction". In: *16th International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 12823. 2021, pp. 55–69.

[9] Denis Coquenet, Clément Chatelain, and Thierry Paquet. "End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 45.1 (2023), pp. 508–524.

References III

[10] Joan Puigcerver. "Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition?" In: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 2017, pp. 67–72.

[11] 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)* (2023).

[12] Emmanuele Grosicki, Matthieu Carré, Jean-Marie Brodin, and Edouard Geoffrois. "Results of the RIMES Evaluation Campaign for Handwritten Mail Processing". In: *10th International Conference on Document Analysis and Recognition (ICDAR)*. 2009, pp. 941–945.

[13] Joan-Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal. "ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset". In: *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2016, pp. 630–635.

[14] Denis Coquenet, Clément Chatelain, and Thierry Paquet. "Faster DAN: Multi-target Queries with Document Positional Encoding for End-to-end Handwritten Document Recognition". In: *International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 14190. 2023, pp. 182–199.