DUBLIN: Visual Document Understanding By using Language-Image Network

Kriti Aggarwal*, Aditi Khandelwal*, **Kumar Tanmay***, Owais Mohammed Khan, Qiang Liu, Monojit Choudhury, Hardik Hansrajbhai Chauhan, Subhojit Som, Vishrav Chaudhary, Saurabh Tiwary

Microsoft Turing

Microsoft

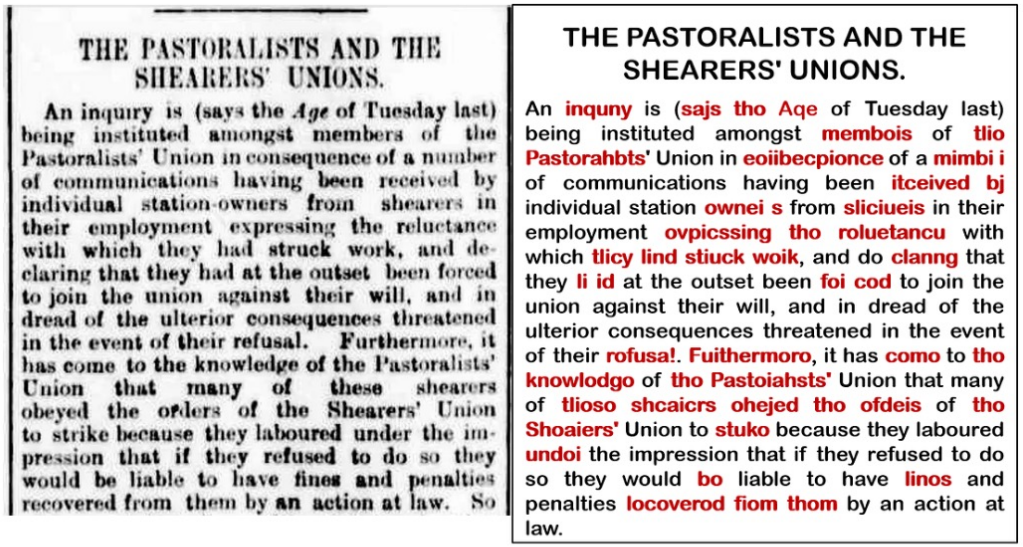
EMNLP 2023

Industry Track

Introduction

Why we need an OCR free model?

- Computationally expensive
- Error-prone Text Recognition
- Inability to handle Rich Visual Cues



THE PASTORALISTS AND THE SHEARERS' UNIONS.

An inquiry is (says the *Age* of Tuesday last) being instituted amongst members of the Pastoralists' Union in consequence of a number of communications having been received by individual station-owners from shearers in their employment expressing the reluctance with which they had struck work, and declaring that they had at the outset been forced to join the union against their will, and in dread of the ulterior consequences threatened in the event of their refusal. Furthermore, it has come to the knowledge of the Pastoralists' Union that many of these shearers obeyed the orders of the Shearers' Union to strike because they laboured under the impression that if they refused to do so they would be liable to have fines and penalties recovered from them by an action at law. So

THE PASTORALISTS AND THE SHEARERS' UNIONS.

An inquny is (sajs tho Aqe of Tuesday last) being instituted amongst membois of tlio Pastorahbts' Union in eoiibecpionce of a mimbi i of communications having been itceived bj individual station ownei s from sliciueis in their employment ovpicssing tho roluetancu with which tlicy lind stiuck woik, and do clanng that they li id at the outset been foi cod to join the union against their will, and in dread of the ulterior consequences threatened in the event of their rofusa!. Fuithermoro, it has como to tho knowlodgo of tho Pastoiahsts' Union that many of tlioso shcaicrs ohejed tho ofdeis of tho Shoaiers' Union to stuko because they laboured undoi the impression that if they refused to do so they would bo liable to have linos and penalties locoverod fiom thom by an action at law.

Figure 1. An excerpt from the OCR-derived text from a newspaper article in Trove (right) and the corresponding scanned image (left). OCR errors are coloured red.

Introduction

- DUBLIN is a Pixel-based OCR-Independent Visual Document Understanding Model
- Pretrained on large number of Webpages and Rendered Images
- Handles diverse tasks like Question-Answering, Information Extraction, Classification, Image Captioning, Machine Reading Comprehension, Bounding box - Text prediction, Natural Language Inference
- Understands and processes various kinds of document images like infographics, charts, forms, tables, natural images, webpages, UI, plain-text
- Achieved SOTA performances by a significant margin (AI2D - 24% ↑, InfographicsVQA - 7.5% ↑, DocVQA - 5.35% ↑)

Date[n 1]			Rank	Tournament name	Venue	City	Winner	Runner-up	Score[1]	Reference
09-09	09-15	THA	WR	Asian Classic	Riverside Montien Hotel	Bangkok	Ronnie O'Sullivan	Brian Morgan	9-8	[2][3]
09-24	09-29	SCO		Scottish Masters	Civic Centre	Motherwell	Peter Ebdon	Alan McManus	9-6	[4]
10-05	10-14	SCO		Benson & Hedges Championship	JP Snooker Centre	Edinburgh	Brian Morgan	Drew Henry	9-8	[5]
10-08	10-13	MLT		Malta Grand Prix	Jerma Palace Hotel	Marsaskala	Nigel Bond	Tony Drago	7-3	[6]
10-16	10-27	ENG	WR	Grand Prix	Bournemouth International Centre	Bournemouth	Mark Williams	Euan Henderson	9-5	[7]
10-29	11-10	THA		World Cup	Amari Watergate Hotel	Bangkok	Scotland	Ireland	10-7	[8]
11-15	12-01	ENG	WR	UK Championship	Guild Hall	Preston	Stephen Hendry	John Higgins	10-9	[9]
12-09	12-15	GER	WR	German Open	NAAFI	Osnabrück	Ronnie O'Sullivan	Alain Robidoux	9-7	[10]
01-02	01-05	ENG		Charity Challenge	International Convention Centre	Birmingham	Stephen Hendry	Ronnie O'Sullivan	9-8	[11]
01-24	02-01	WAL	WR	Welsh Open	Newport Leisure Centre	Newport	Stephen Hendry	Mark King	9-2	[12]
02-02	02-09	ENG		Masters	Wembley Conference Centre	London	Steve Davis	Ronnie O'Sullivan	10-8	[13][14]
02-13	02-22	SCO	WR	International Open	A.E.C.C.	Aberdeen	Stephen Hendry	Tony Drago	9-1	[15][16]
02-23	03-02	MLT	WR	European Open	Mediterranean Conference Centre	Valletta	John Higgins	John Parrott	9-5	[17][18]
03-10	03-16	THA	WR	Thailand Open	Century Park Hotel	Bangkok	Peter Ebdon	Nigel Bond	9-7	[19][20][21]
03-18	03-23	IRL		Irish Masters	Goff's	Kill	Stephen Hendry	Darren Morgan	9-8	[22][23]
03-27	04-05	ENG	WR	British Open	Plymouth Pavilions	Plymouth	Mark Williams	Stephen Hendry	9-2	[24]
04-19	05-05	ENG	WR	World Snooker Championship	Crucible Theatre	Sheffield	Ken Doherty	Stephen Hendry	18-12	[25]
05-??	05-??	WAL		Pontins Professional	Pontins	Prestatyn	Martin Clark	Andy Hicks	9-7	[26]
12-28	05-18	ENG		European League	Diamond Centre	Irthlingborough	Ronnie O'Sullivan	Stephen Hendry	10-8	[27]

Question: What is the name of the first venue on this list?
DUBLIN's Answer: Riverside Montien Hotel
Gold Answer: Riverside Montien Hotel

Model Pretraining Framework

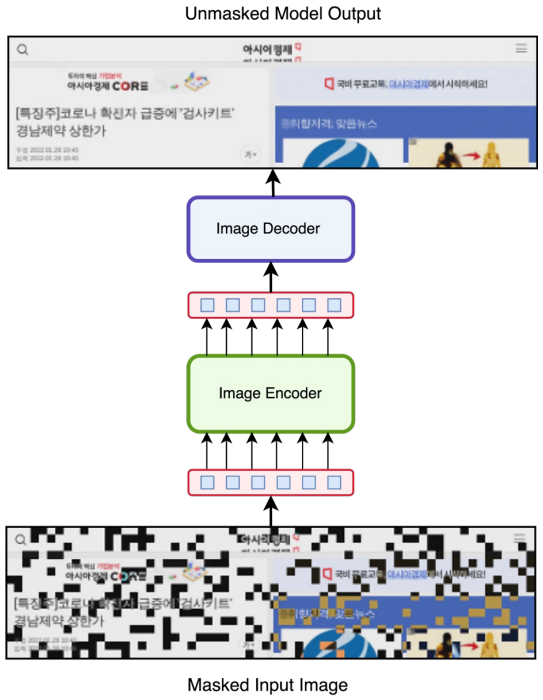


Figure 1

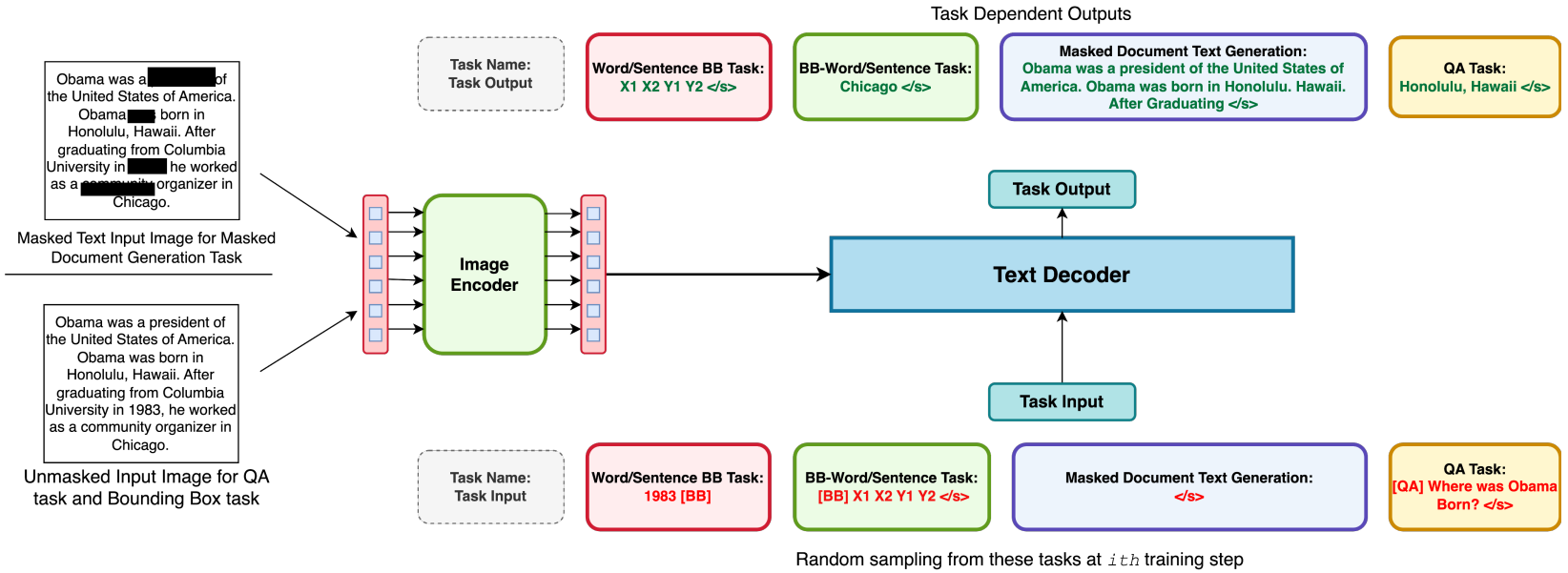
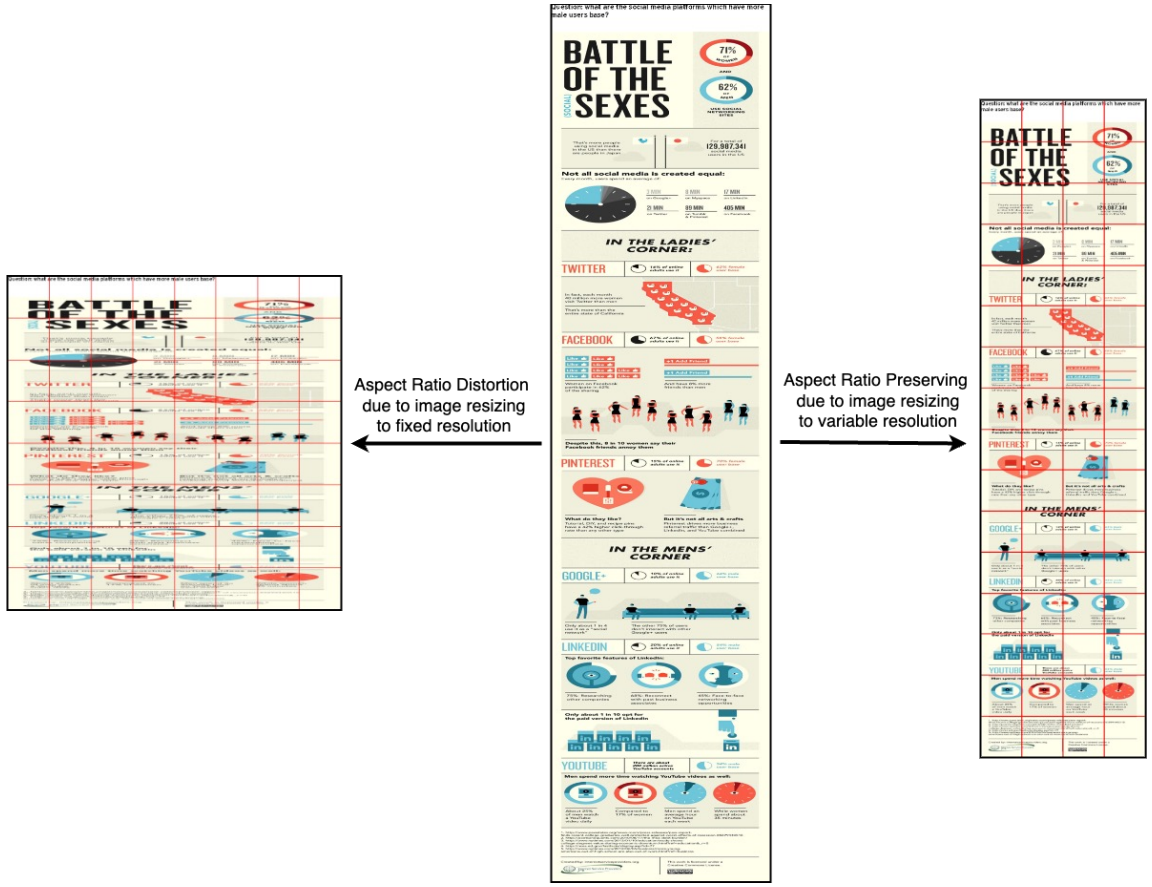


Figure 2

Pre-processing Techniques before Finetuning

- Question on top of the images
- Variable Input Resolution to handle documents of different aspect ratios
- Template based finetuning to execute different kinds of tasks without adding any external layers for specific tasks.



Results

Model	QA over Illustrations			UI understanding			Captioning	Document QA	
	ChQA	**AI2D**	**O-VQA**	**RefExp**	**Widget Cap**	**Scrn2Wds**	**TCaps**	**DVQA**	**IVQA**
Metrics	RA	ANLS	F1	EM	CIDEr	CIDEr	CIDEr	ANLS	ANLS
Donut	41.8	30.8	66.0	-	127.4	56.4	74.4	67.5	11.6
$\text{Pix2Struct}_{\text{large}}$	**58.6**	42.1	71.3	94.2	**136.7**	**109.4**	**95.5**	76.6	40.0
$\text{Dublin}_{\text{fixed\_res}}$	35.6	**51.1**	**73.1**	**99.1**	132.2	101.8	92.8	**78.2**	**36.8**
$\text{Dublin}_{\text{variable\_res}}$	35.2	**52.3**	**74.0**	**99.1**	132.2	101.8	92.8	**80.7**	**43.0**
(SOTA with) spl. pipelines	(VTP) 45.5	(DQAN) 38.5	(LATr) 67.5	(UIB) 90.8	(VUT) 97.0	(VUT) 64.3	(PaLI) 160.4	(UDOP) 84.7	(UDOP) 47.4

Table 1: Performance on QA over illustrations, UI understanding, image captioning and QA tasks. Higher the better. ChQA: ChartQA, O-VQA: OCR-VQA, Scrn2Wds: Screen2Words, TCaps: Text Captioning, DVQA: DocVQA, IVQA: InfoVQA, VTP: Vision Tapas Model (Masry et al., 2022), DQAN: Diagram Question-Answering Network (Kembhavi et al., 2016), LATr: Layout-Aware Transformer for Scene-Text VQA (Biten et al., 2021), UIB: UI-Bert (Bai et al., 2021), VUT: Versatile UI Transformer (Li et al., 2021b), PaLI: Pathways Language and Image model (Chen et al., 2023).

Results

Model	Information Extraction			Classification	Reading Comprehension	
	FUNSD	**CORD**	**DeepForm**	**RVL-CDIP**	**WebSRC**	**VisualMRC**
Metrics	F1	F1	F1	Accuracy	EM/F1	CIDEr
Donut	-	91.6	-	**95.3**	-	-
$\text{Dublin}_{\text{fixed\_res}}$	**77.8**	**97.1**	62.2	94.9	**77.7/84.2**	**347.3**
$\text{Dublin}_{\text{variable\_res}}$	**77.8**	**97.1**	**65.7**	94.9	**77.7/84.2**	**347.3**
SOTA with Spl. Pipelines	(LyLMv3) 92.08	(UDOP) 97.6	(UDOP) 85.5	(UDOP) 96.00	(TIE) 81.6/86.2	(LyT5-large) 344.1
$\text{BERT}_{\text{large}}$/T5 (Text Baseline)	65.63	90.25	74.4	89.92	-	-

Table 2: Performance on IE, doc classification, WebSRC and VisualMRC. Higher the better. LyLMv3: LayoutLMv3 (Huang et al., 2022), LyT5-large: LayoutT5-large (Kembhavi et al., 2016).

Model	Table QA/NLI		
	WTQ	**TabFact**	**WikiSQL**
Metrics	EM	Accuracy	EM
$\text{Dublin}_{\text{fixed\_res}}$	25.7	**73.54**	75.3
$\text{Dublin}_{\text{variable\_res}}$	**29.7**	72.9	75.3
(SOTA w/) Spl. pipelines	(UDOP) 47.2	(UDOP) 78.9	(TAPEX) **89.2**
(BART) Text Baseline	38.0	76.0	85.8

Table 3: Performance on Table QA and NLI. Higher the better.

Examples

Rene Kok

Visitor Experience

Test Engineer | Europeana Foundation | Netherlands | GLAM

rene.kok@europeana.eu

Rene wants people to know that being a test engineer is one of the most misunderstood jobs on earth. Most people think a test engineer tries to find all bugs and errors in software. This is not the job of a test engineer. A test engineer tries to find out if the software is good enough to let the users play with it. So in reality, he does hunt bugs and errors while determining if the software is ready for the users. To do so, test engineers practice voodoo that allows them to get a decent understanding of the quality of the software with a minimum of testing (so test engineers are masters at cutting corners to find out what they need to know).

Question: What does Rene want people to know about being a test engineer?
DUBLIN's Answer: **He wants people to know that being a test engineer is one of the most misunderstood jobs on earth.**
Gold Answer: That being a test engineer is one of the most misunderstood jobs on earth.

Modou Bamba Gaye

From Wikipedia, the free encyclopedia

Modou Bamba Gaye is a Gambian politician who was the National Assembly Member for Lower Saloum, representing the National Reconciliation Party (NRP), from a 2015 by-election to the 2017 parliamentary election.

Political career [edit]

Gaye was elected at a 2015 by-election for the seat of Lower Saloum, following the dismissal of incumbent NAM Pa Malick Ceesay from the ruling Alliance for Patriotic Reorientation and Construction (APRC). Gaye defeated APRC candidate Kebba Touray in the election, winning 2764 votes to Touray's 1618 votes.[1] Speaking in the National Assembly in January 2017, during the constitutional crisis and Yahya Jammeh's refusal to step down, Gaye called for a peaceful transition of power and said, "The people who voted us in are the same people who voted for Jammeh before and are the same people who voted Adama Barrow."[2]

Question: When was Gaye elected for the seat of Lower Saloum?
DUBLIN's Answer: **Gaye was elected at a 2015 by-election.**
Gold Answer: In 2015

Conclusion

- DUBLIN is a 976M parameter model which can handle diverse types of document images and perform different kinds of task.
- DUBLIN is a versatile and robust model that does not rely on external OCR systems and can be finetuned in an end-to-end fashion.
- We also introduce a new evaluation setup on text-based datasets by rendering them as images.
- This model can be used in various applications, from search engines to presentations.
- Possible future direction: Integrating Generative models like T-NLG or Llama models.