Language Modelling from Pixels



Desmond Elliott

Language and Multimodal Processing Group Department of Computer Science University of Copenhagen

Warning: The final part of the talk contains dataset samples that are racist in nature.

What is Language Modelling?

• A language model is a statistical model of text and can be used to estimate the probability of a string **x** consisting of *T* tokens.



Tractable with assumptions, e.g. bi-gram model

Modern Language Models



Okay but why is it useful?

- Spelling and grammar checking
- Machine translation
- Web search
- Text prediction
- Topic modelling
- Chatbots



Current Status: Scale



Treviso et al. 2023. Efficient Methods for Natural Language Processing: A Survey. TACL Wei et al. 2022. Emergent Abilities of Large Language Models. TMLR

NLP for All Written Languages

- There are 3,000 written languages
 400 with >1M speakers
- NLP usually covers 100 languages
 Technological exclusion for billions





NLP is a pipeline ...



Syntactic / Semantic analysis

... that is easily broken



This issue disproportionately affects low-resource languages

The Vocabulary Bottleneck

- NLP is an **open vocabulary problem** and the ability of a model is determined by its vocabulary:
 - 1. tokens, characters, sub-words, etc.
- This creates a bottleneck in two places:
 - 1. Representational bottleneck in the Embedding layer
 - 2. Computational bottleneck in the Output layer



Where's the sweet spot?

Long sequence lengths

(CANINE, CharFormer) Characters

(RoBERTa: 50K) Medium subword vocab

(XML-R: 250K) Large subword vocab

Bytes (ByT5)

Small subword vocab (BERT: 30K)

Full unicode coverage (ByT5, CANINE)

Words (word2vec: 3M)

Large vocabulary

Treat language as vision





Overview

1. Pixel Language Model
 2. Text Rendering Matters
 3. Historical Document Processing

The Model



Flexible Text Renderer

My cat 🦮 loves pancakes 🥮 and my duck 🦆 loves grapes 🍓.

- Emoji
- 🕒 🕒 🕪 اتنشط القطط في الخلاء ليلا ونهارا على الرغم من أنها تميل إلى أن تكون أكثر نشاطا بقليل في الليل. 🚽 L 🖉 他們常在晚間活動,但並不表示他們是夜行性動物。

• Wc ድመት በአሁኑ ጊዜ ከሁሉም እንስሳ በላይ በቤት እንስሳነቱዋ ተፈላጊነትን ያላት ናት ።

A new type of generative model



Pretraining

- English Dataset: English Wikipedia and Books Corpus
- Masking: 25% Span Masking
- Maximum sequence length: 529 patches (16x8464 pixels)
- Compute: 8 x 40GB A100 GPUs for 8 days
- **Parameters**: 86M encoder + 26M decoder

There is only 0.05% non-English text in our pretraining data (estimated by Blevins and Zettlemoyer 2022)

The Great Wall of China (traditional Chinese: 萬里長城; simplified Chinese: 万里长城; pinyin: Wànlǐ Chángchéng)

Downstream Tasks

• Datasets: Universal Dependencies, MasakhaNER, GLUE, Zeroé

• Models:

	Parameters	Pretraining Data	
PIXEL _{BASE}	86M	English Wikipedia + Bookcorpus	
$BERT_{BASE}$	110M	—	Similar pretraining setup
CANINE-C	127M	104-languages from Wikipedia	Tries to solve the same problem using UTF-32

Dependency Parsing Results





PIXEL vastly outperforms BERT on unseen scripts

Named Entity Recognition in African Languages



BERT 🔳 CANINE 🔳 PIXEL



PIXEL outperforms BERT on the non-Latin script

μ

PIXEL outperforms the multilingually pretrained CANINE-C

20

GLUE: Sentence-level Understanding

BERT PIXEL 100 75 50 25 0 MNLI QNLI SST-2 COLA STS-B MRPC RTE QQP WNLI BERT outperforms PIXEL on English sentence-level tasks

Score

Adapting to new fonts





2. Text Rendering Matters

Text Rendering Matters

Our original text renderer produces many nearly-identical patches
 This is representation- and compute-wasteful



Can we do better?

Lotz et al. Text Rendering Strategies for Pixel Language Models. EMNLP 2023.

Alternative Text Renderers



GLUE (revisited)

BERT PIXEL PIXEL-BIGRAM



Bigram text rendering produces better models

Scaling Down

• Better text rendering can create effective models at smaller scales



3. Historical Document Processing

Warning: This part of the talk contains dataset samples that are racist in nature.

Historical Document Processing

- Worldwide efforts to digitize historic documents (Groesen 2015)
- Typical pipeline for enabling access is:
 - a. Scan documents into high-quality digital formats
 - b. Perform OCR on those documents (one-off process)
 - c. Search through documents using OCR annotations

What if we could do this without OCR?

Caribbean Newspapers, 1718–1876

- Collaboration with researchers that are interested in tracking newspapers notices about escaped slaves
 - What was the given name?
 - What reward was offered?
 - Who was the contact person?
- Dataset of 1.65M scanned pages

BORNI SERVICE CONTRACTOR					
0. 24Vol. XXXIX.	24s. per Ann				
	Hamilton, Bermuda, 7	uesday, June 19, 1866.			
Commissariat, Bermuda.	BERMUDA, Alies	Mechanics' Industrial			
HABILTON, ITYE JUNE, 1806.	By His Excellency HARRY ST. GEORGE	TEELSITION,	200 B		
DEPUTY COMMISSION OF DEPUTY COMMISSION OF	enois O der of the Roll, Brief.	In aid of Completing the Association's	CARE OF THE OWNER		
s in Humilton, until Noon of	Geboul in the Rugal Engineers, Gebouar, Commander in Ohl f, Vite	Hall.	SOMERS' ISLANDS.		
The 23rd June.	Admind and Ordinary, is and our three (stands, Sec., Soc., Sec.,	HIS EXCELLENCY THE GOVERNOR	By His Estellency HARRY ST. GEORGE OPD Com		
an Persons willing to supply such Quantities of	WBEREAS MARY FRANCES PITCHER	AND NIIS ORD.	[L.S.] punion of the Most Har-		
BUDG	amened, on the Entire of Althout ULARK	Wednesday 27th, Thursday 28th, and	H. St. George Ord, Breat-Colouri in the Nayal		
	Georges Parish in three felaude, Hiomenianon, De-	Friday 28h June, on the Property of Mas, KENNEDT, KEOPE 28	der-in-Chief. andr-in-Chief and Vice-		
decises, between the lot Joly, 1806 and 31st	This is therefore to give Notice, that if any Person	"Richmond Grounds."	Admiral in and over these Islands, he., he., he.		
anterly. Further information can be obtained	or Persons can show any jost Canae why the said Administration should not be reasoned from the said		A Proclamation.		
apploying at the Consistent Orpors at Gauger.	MART PRANUES Privars, Inc. she, or they are to	A EXCELLENCY THE GOVERNOR has	Governon and Constantion has reached Mn. The		
T. W. GOLDIE.	Secretary's Office of these falends within Fifteen	NARLY CONSISTENCES IN A DATE OF THE STATE OF	Perts of HALIPAX and NEW YOREs-1 DO		
D. C. O.	Administration will be granted scentilingly.	THORESONT.	THEREPORT. by victor of the power and natio-		
[Maximum papers losert twice.]	NILES GERALD KEON,	27th June,	these balands, intituded, "As Art to consolidate and		
rticles Adapted to the wants of all	Dated at the Secretary's Office, }	viz, 2-the 25th and 20th.	advice and Consent of Hat Nations's Control.		
Classes of Society,	For Solo	is not some some init that, when completed, is to rand shirtly for edicard mulphrpo-es, which is haped	CLANATION, and do hereoy rule known that		
On application at the	RY "FLIZA RAPSS"	to prove advisit agrees to the Country at large, the Committee most earnesity solirs oil by uns of	ed Places within the meaning of the said, bet And		
t. Georges, General Store,	AND IN STORE.	derutions from every class of the Pable. Contribu-	I do neersby strictly charge and Command all Pilots going on board or taking charge of any wanel gener-		
	CARBELS Nov T. N. PORK.	revelved at the Store of Mn. Hanneyr, Hamilton,	ing at these Islands from either of the afterestill Ports forthwith to conduct the supe to approximate it		
In addition to other second Arrients,	D Dino divo Packet Mess BEF.P	to ef for the purpose through the courtory of Man-	Quarantine Platicus pretorical by the above named		
ARE RECEIVING	Barrels FLOUS and CORN MEAL	During the time of the Exhibition every cure will	HAALTH OFFICER, who shall thereapon give much		
er ' Minnie Ha Ha,' ' Forest Fairy.'	Eags White and Yellow CORN Eags 5.8.AN, 5 bushels each	be taken to promote the constort of the visitors. Neo'do will be erected for shelter from the nan, and	ense may justify and to his said office may pertain.		
"Star of the East,' &c-	Barrels Brown and White SUGAR Boxes Honry Des TUBACCO, 1276	Refrestments in grant variety will be perpared for the occusion, which, in co-junction with the display	Giten u der My Marst and the Great Fest of these Islands his Swand		
The following :	Brietol's SURSUPARILLA, SO VP and STARCH HAVE and EACOE	of Goods, both lassi and f corry-hidsorio unoqual-	day of May, 1008, and in the		
Shoemek-rs' TOOLS,	Administre and Tallow CANDLES	being made, it is hoped that all who may visit the	jesty's Reign.		
avering PUTS Tuller CANS	tuna Demorara r.UM, deo, dec, Sec.	By authority of the Consultor,	MILES GERALD KEON.		
in Tes KETTLES Cast Iron DO., (tinned)	B. W. WALKER.	C. W. GAUNTLETT,	Colonial Secretary		
Dranined Round and Ovel TUNS	0.	May 29, 1806. Secretary.	Gen SAVE VIE QUEEN		
IN POTE Table BF.LLS TOAM FORKS IR PANS Coffice MILL ^A Sugar SCUOPS CALE ⁴ and WEIGHTS, § or. to 40., to 185a.	Offers for Sale,	ICE.	.0.8.7.		
to 56h, and to 200 lbs. Iso, Spring BALANCES	Ex barque Eliza Barss,	THE SUBSCRIBERS	2000		
undrant DITTO, &co., Coffee MILLS opped, Ovel and Equare Bake PANS	FROM NEW YORK,	Ara New Ropeiring	BERMUDA, Alles,		
upan'd CANS Remains JACKS	10 Barrels TAR.	THEIR USUAL SUPPLY OF	By His Exclinity HARRY ST. GEORGE		
ownelled SAUCEPANS	mup or poop pp	ECE	ORD, Comparison of the Next Honographic Order		
KNIVES, Pitole SCHSORS &-	THE SUBSCRIBER	Which they will competers to have	[L. S.] nel in the Brazi Envi		
an PAPER Emery CLOTH	His usual summary of	On the 1st June.	H. St. George Ord, Marris, Gonerman, Com		
TOONS and HANDLES	SUMMER GOODS	TERMS will be made known on opplication at	mander-in-Chief. Vice-tduiral in an		
ex, Danster and Choo BRUSHES	FROM LONDON.	GOSLING BROS.	he., he.		
olat SETS, Acc., 150 dozen BASINS-mit-	per Mull Stowner via H-Kfax,	Dawline, May 50, 1898.	A proclamation.		
od to Military and Noval Messee, it and Pr-st GLASS WINES TUMBLERS	Which he offers at a small advance for Cash at his Residence.	AICLOBIT WOLFF	W Governa And Coveration of Children and		
RESERVES Creet BOTTLES SALTS avar BASENS Butter DISHES	FOSTER L. BONNELL	Front Street, Hamilton.	LOUPE, one of the Fretch West India Islands ;-		
ik EWERS, &co., &co.	Bannens Bay, June 4m, 1999 -	br in farmer Proprieters, Nes. C. NANTER,	 DO THEREFURE, by virtue of the power a data thority is no vested, by an Act of the Lagislature of 		
23" IN.B. Harnesses, Sc., neally Made	TEAS and COFFEE.	who miniful at former liberal pstrougge extended to her, and feeling granuful for all part fevors again	thus labed, intituded "An Act to Comulate and Amend the Querastite Lows," and by and with		
OXBORROW & HUGHES.	TTALP Cheets Congon TEA.	vectores to selicit the support of her Priezds and the Pablic generally in the revived E-tablichment	the advice and consent of Hos Marsorr's Cornell		
k. Georges, Juan, 9th 1863.	Half Ditto Southing DITTO, Half Ditto Orlean DITTO.	which the trasts will exciting to desorve and re-	MATION, and do horeby make knows that the		
For Sale.	Half Ukto but Byen DITTO,	BETARFANTS, LENGEROUS, DIXXER+, TEAS, &c.	within the meaning of the axid Act		
Per Recent Importations.	Media COPPEB,	Terme.	pring on board or taking charge of any Vessel arrive		
And per ELIZA BARSS,	Java DITTO.	The trease is new ready to receive Hoarders. Hamilton, April 07th, 1804,	and at mean labored from the events of the dest		
Just from New York,	Wholesale or Retail,	SODA WATER	ansine Stations prescribed by the above named A et there to remain want she whall be visited by the		
Ditte Pilot BREAD, avail cakes	CONTING PROTEINS	Soba wallsk.	BRALTS OFFICEL who shall thanspas give such		
SIRES CHUCOLATE Roxes HERRINGS	Hamilton and St. Georgee.	Bottled Soda Water.	ease may justify, and to his said office may parts of		
Choice BUTTER and CHEESE.	October 23, 1813.	Of a Superior quality can be sup-	Given unfer My Hand and the 44-m Seal of these laberds rue		
A REAL PROPERTY AND A REAL		a set of the set			

PHD: PIXEL for Historical Documents

- Historical document-aware Pretraining
 - Mixture of scanned newspapers and synthetic newspaper-like text generated from Wikipedia and Bookcorpus datasets
 - All input data is scaled to 368x368 and split into 16x16 patches

sionally blogs such as Arcade, a humanities site published by Stanford University. From 2012 to 2016, he hosted a radio show webcast by Alanna Heiss's Clocktower Productions. In autumn 2020, an article he wrote for The Creative Independent was widely disseminated on the internet. Called 19 things I'd tell people contemplating starting a record label (after running one for 19 years) it was a mix of advice, warnings, and personal history gleaned from almost two decades of operating Brassland. It was followed by an appearance on the Third Story podcast.

Sickman's war service took him to Tokyo during the occupation of Japan where he served as one of the "Monuments Men" under General Davidas MarArthur's terministical by the Hit Englished club in 1991 in ander for The Championships, Wimbledon to be held Since then the club has been normadic, moving to Distarley and Greenford before settling in Acton and playing their matches at Juspes FC's Twyford Avenue Sports Ground By 2024, the club had downsized to running only one team. A number of players for the New Zealand

national rugby urion team have played for London New Zealand Including Daug Rollerson, Tenry Marrison and Paul Scaeford In recognition of their history, the club have been granited privileges from both the Rugby They are the only rugby team aside of New Zealand national representative teams that wears the silver fern as their creat and the RPU exempted them from the oversees player quotes, prior to their abilition. The club have alise failed part in a number of New Zealand aving been estranged from her father's family for most of her life, Andrea is intrigued. But what exactly is the Bancroft's involvement with "Genesis," a mysterious person working to destabilize the geopolitical balance at the risk of millions of lives? In a series of devastating coincidences, Andrea and Belknap come together and must form an uneasy alliance if they are to uncover the truth behind "Genesis"—before it is too late.

Girls' BMX was part of the cycling at the 2010 Summer Youth Olympics program. The event consisted of a seeding round, then elimination rounds where after three rores the ton 4 achieved qualifying standards in the following events (up to a maximum of 2 swimmers in each event at the Olympic Qualifying Time (OQT), and potentially 1 at the Olympic Selection Time (OST)): Venezuela has entered one athlete into the table tennis competition at the Games. Gremlins Arvelo secured the Glympic spot in the women's singles by virtue of her top six finish at the 2016 Latin American Qualification fourmament in Santiago, Chile.

Visual Question Answering in Newspapers

• Frame this as a Visual Question Answering Task



Results



Surprisingly good performance compared to a model trained on manually transcribed text

Conclusions

- PIXEL is a new type of language model that tackles the open vocabulary problem using visually rendered text
 - 1. This enables high-quality transfer to different scripts
 - New, unseen languages
 - Different fonts in existing languages
 - 2. Compact models with as few as 5M parameters
 - 3. Natural interface to scanned documents

Open Questions

Thanks













- N. Borenstein
- P. Rust
- I. Augenstein

E. Bugliarello

E. Salesky

M. de Lhoneux







Code

Models